Spring 2007

Canonical Correlation and Correspondence Analysis of Longitudinal Data

Jayesh Srivastava
Old Dominion University

CANONICAL CORRELATION AND CORRESPONDENCE ANALYSIS OF LONGITUDINAL DATA

by

Jayesh Srivastava
M.S. August 1998, Indian Institute of Technology

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirement for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTATIONAL AND APPLIED MATHEMATICS

OLD DOMINION UNIVERSITY
May 2007

Approved by:

Dayanand Naik (Director)

N. Rao Chaganty (Member)

Larry Lee (Member)

Edward Markowski (Member)

ABSTRACT

CANONICAL CORRELATION AND CORRESPONDENCE ANALYSIS OF LONGITUDINAL DATA

Jayesh Srivastava
Old Dominion University, 2007
Director: Dr. Dayanand Naik

Assessing the relationship between two sets of multivariate vectors is an important problem in statistics. Canonical correlation coefficients are used to study these relationships. Canonical correlation analysis (CCA) is a general multivariate method that is mainly used to study relationships when both sets of variables are quantitative. When the variables are qualitative (categorical), a technique called correspondence analysis (CA) is used. Canonical correspondence analysis (CCPA) is used to deal with the case when one set of variables is categorical and the other set is quantitative. By exploiting the interrelationships between these three techniques we first provide a theoretical basis for CCPA.

Next, in this dissertation, we have generalized each of these three techniques to analyze the relationships between two sets of repeatedly or longitudinally observed data. When the two vectors are quantitative, we use a block Kronecker product matrix to model dependency of the variables over time. We then apply canonical correlation analysis on this matrix to obtain canonical correlations and canonical variables. When the variables are qualitative, the data are summarized in the form of a contingency table. It is generally not straightforward to model dependency of contingency tables over time. However, we have proposed fitting correlated linear models to the summary statistics obtained by performing the usual correspondence analysis at each time period. We have shown that the most useful summary measure for this purpose is the first singular value of the correspondence matrix, which is essentially the matrix of relative frequencies obtained from the given contingency table. Our method is a reasonable approach to analyze repeated contingency table data. Finally, to deal with the case when one set of variables is categorical and

the other set is quantitative, we have proposed combining the two approaches to deal with quantitative and qualitative variables. We have illustrated and studied the performances of our methods my implementing them on simulated data sets.

High dimensional data are now common due to the Internet, genomics, proteomics, and the like. Although, correspondence analysis and other methods considered in this dissertation are general techniques for analyzing multivariate data their usefulness for analyzing very high dimensional data have not been compared with the other more modern machine learning methods. In the last chapter of this dissertation, we provide a brief introduction to a machine learning method that is used to analyze very high dimensional and sparse contingency table data from the field of language processing or information retrieval, named latent semantic analysis (LSA). We then propose certain criteria to compare the performance of LSA with the correspondence analysis. Based on these criteria we find that under certain situations correspondence analysis performs better.

ACKNOWLEDGMENTS

This dissertation could not have been completed without the invaluable help of a large number of individuals to whom I am gratefully indebted. Dr. Dayanand Naik read countless drafts and revisions, and provided guidance and help at every stage. He shaped my thoughts on multivariate statistics in myriad ways and constantly encouraged me to be a focused researcher, exploring current theories in canonical correspondence analysis. Similarly, Dr. N. Rao Chaganty left a lasting impression on my intellectual and personal development. His standard of excellence was second to none, and he continually pushed me to be better than what I was or could be. Without his mentoring and guidance, early on, such an undertaking would not have been completed.

I would like to thank Drs. Larry Lee and Edward Markowski for serving on my dissertation committee. I also like to thank Dr. Larry Lee for editorial help. Thanks are also due to Dr. Irwin Levinstein for his immense encouragement and financial support. Special thanks to Dr. Deepak Mav for providing me an efficient SAS code to simulate multivariate Poisson random variables. I would also like to thank the faculty and staff in the department of Mathematics and Statistics at Old Dominion University. In particular, Dr. Ram C. Dahiya, Dr. John M. Dorrepaal, Dr Hideaki Kaneko, Barbara Jeffrey and Gayle Tarkelsen who have shared with me infinitely more patience and wisdom than I ever deserved.

Finally, I would like to thank all my colleagues and friends, who provided me their support and put up with me in this phase of my life.

I dedicate this thesis to my parents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The main focus of this dissertation is to provide methods to study the relationships between two sets of repeatedly or longitudinally observed data. Methods are developed for all the three different cases, namely, when (i) both sets of variables are quantitative, (ii) both sets are qualitative, and when (iii) one set is quantitative and the other set is qualitative. These cases are considered in separate chapters that follow. A brief introduction to each chapter is provided next.

Studying the relationship between two sets of variables is an important multivariate statistical analysis problem. Hotelling (1936) introduced his famous canonical correlation analysis (CCA) to study the relationship between two sets of quantitative variables. In this analysis, one finds a linear combination of the first set of variables and a linear combination of the second set of variables such that they both have unit variance and the Pearson correlation coefficient between them is maximum. Thus the obtained pair of linear combinations are called the first canonical variables and the correlation is called the first canonical correlation. This process is repeated to obtain the second, third,... canonical variables and correlations with the additional restriction that the pair of linear combinations currently being computed are uncorrelated with all the previously obtained pairs. Use of few canonical variables to perform data analysis is in fact a general way of dimension reduction. Although CCA has been generalized in several directions (see Kettenring (1971)), its generalization to deal with longitudinally observed sets of variables has not been done in the literature. In the next chapter (Chapter 2) we provide canonical correlation analysis of longitudinally observed sets of data. Suppose two random vectors $\mathbf{x}$ and $\mathbf{y}$, of dimensions $p \times 1$ and $q \times 1$ respectively are observed over t time periods on n subjects. Then assuming a block Kronecker product variance covariance matrix to the $(pt + qt) \times 1$ random vector we account for the dependency of the variables over time. Various testing of hypothesis problems under this scenario are considered and the CCA using these matrices is illustrated on simulated data sets.

When both variables are qualitative, the data are summarized in the form of a

This dissertation follows the style of *Journal of the American Statistical Association.*

contingency table. Correspondence analysis (CA) is a method that is used to project the relationship between the two qualitative variables on to a smaller dimension space. Benzécri (1969, 1992), Hill (1974), Greenacre (1984), and others have studied correspondence analysis in detail and provided numerous applications. Also see Khattree and Naik (2000). Correspondence analysis results can be obtained by performing canonical correlation analysis on certain matrices. However, generalization of the CCA approach proposed in Chapter 2 of this thesis to deal with repeated data does not apply to repeatedly observed contingency tables. The reason why the CCA approach fails here is because it is difficult to model the dependency of contingency tables over time by the usual correlations. To overcome this problem, in Chapter 3 we have proposed fitting generalized linear models to the summary statistics of CCA corresponding to each time period. This gives us a reasonable approach to handle analysis of repeated contingency table data. Illustration of the methods is performed on simulated data. Assuming the frequencies in contingency tables are independently distributed as Poisson, we use certain extensions of an algorithm due to Sim (1993) to generate correlated Poisson frequencies over time periods. These extensions of the algorithm and the SAS code implementing the algorithm are given in Mav (2004) and Chaganty and Mav (2007).

Canonical correspondence analysis (CCPA) is used to deal with the case when one variable is categorical and the other set of variables is quantitative. The method was introduced by Ter Braak (1986) to analyze species abundance and environmental variables data obtained at a certain number of sites. In Chapter 4 we review this method and show that the results obtained using CCPA too can be obtained by performing CCA on a set of matrices obtained from the data. In the literature, we found no population versions to these matrices. Using the approach of Olkin and Tate (1961) we provide a theoretical basis to CCPA in Section IV.3. Then we propose methods to deal with repeated data by combining the approaches that we have taken in Chapters 2 and 3. Methods are illustrated using simulated data sets.

High dimensional data are now common due to the Internet, genomics, proteomics, and the like. Although, correspondence analysis and other methods considered in this dissertation are general techniques for analyzing multivariate data their usefulness for analyzing very high dimensional data have not been compared with the other more modern machine learning methods. In Chapter 5 of this dissertation,

we provide a brief introduction to a machine learning method that is used to analyze very high dimensional and sparse contingency table data from the field of language processing or information retrieval, named latent semantic analysis (LSA) (Deerwester et al., 1990). We then propose certain criteria to compare the performance of LSA with the correspondence analysis. Based on these criteria we find that under certain situations correspondence analysis performs better.

Most of the computations and simulations are done using IML procedure in SAS software. The results from different chapters are provided in numerous tables and figures.

CHAPTER II

REPEATED CANONICAL CORRELATION ANALYSIS

II.1 Introduction

Canonical correlation analysis (CCA) is a well known statistical technique used to identify and measure the association between two sets of random vectors using specific matrix functions of variance-covariance matrices of these variables. This is also one of the most general methods for data reduction in multivariate analysis. CCA was introduced by Hotelling (1936) while studying the relationship between two sets of variables in instructional research. Now CCA has found many applications in different fields and it is routinely discussed in many multivariate statistical analysis textbook. For example, see Mardia, Kent and Bibby (1979) or Johnson and Wichern (2002). Suppose the random vector $\mathbf{x}$ of p components and random vector $\mathbf{y}$ of q components have the variance-covariance matrix $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{yy}$ respectively and suppose $\mathbf{\Sigma}_{xy} = \text{cov}(\mathbf{x}, \mathbf{y})$. That is,

$$D\begin{bmatrix}\mathbf{y}\\ \mathbf{x}\end{bmatrix} = \begin{bmatrix}\Sigma_{yy} & \Sigma_{yx}\\ \Sigma_{xy} & \Sigma_{xx}\end{bmatrix}.$$

The main idea behind canonical correlation analysis is to find a $q \times 1$ vector $\mathbf{a}$ and a $p \times 1$ vector $\mathbf{b}$, given $\mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{yy}$ and $\mathbf{\Sigma}_{xy}$, so that the correlation between $\mathbf{a}'\mathbf{y}$ and $\mathbf{b}'\mathbf{x}$ is maximized.

The i^{th} pair of canonical variables $(\mathbf{a}_i'\mathbf{y}, \mathbf{b}_i'\mathbf{x})$ is obtained by solving

$$\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\mathbf{a}_i = \rho_i^2\mathbf{a}_i \qquad (2.1.1)$$

$$\text{and } \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\mathbf{b}_i = \rho_i^2\mathbf{b}_i, \qquad (2.1.2)$$

where ρ_i is a canonical correlation and ρ_i^2 is eigenvalue of $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1/2}$.

Kettenring (1971) has generalized CCA to several sets of variables and it has found many generalizations in the literature. Beaghen (1997) has used canonical variate method to analyze the means of longitudinal data. However, no methods have been developed to perform CCA on longitudinally observed data. Focus in this chapter is to generalize canonical correlation analysis to repeatedly observed data on $\mathbf{x} = (x_1, ..., x_p)'$ and $\mathbf{y} = (y_1, ..., y_q)'$.

II.2 Repeated Measures Case

Suppose we have observed $\mathbf{x}$ and $\mathbf{y}$ repeatedly over t time periods. Let $\mathbf{x}_i$ and $\mathbf{y}_i$ be the vectors $\mathbf{y}$ and $\mathbf{x}$ observed at the i^{th} occasion. Define $\mathcal{Y} = (\mathbf{y}_1', \ldots, \mathbf{y}_t')'$ and $\mathcal{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_t')'$.

covariance matrix $\mathbf{D}$ of $\mathbf{u} = (\mathcal{Y}', \mathcal{X}')'$ has a Kronecker product matrix structure. coThat is,

T

$$\mathbf{D} = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix}. \tag{2.2.1}$$

$$\mathbf{D} = \begin{bmatrix} \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix}. \tag{2.2.1}$$

The matrices Ω_{yy}, Ω_{xx} and Ω_{yx} are used to model the dependency over t time period of repeated measurements on $\mathbf{y}$, on $\mathbf{x}$ and of the covariance matrix between repeated measures of $\mathbf{y}$ and $\mathbf{x}$ respectively. Kronecker product structures have been successfully utilized to analyze multivariate repeated measures data in Naik and Rao (2001) and Chaganty and Naik (2002).

The problem here is to determine linear functions $U = \mathbf{a}'\mathcal{Y}$ and $V = \mathbf{b}'\mathcal{X}$ such that the correlation between them is maximum. Here $\mathbf{a}$ is $qt \times 1$ and $\mathbf{b}$ is $pt \times 1$ vectors. Assuming $E(\mathcal{Y}) = 0$, $E(\mathcal{X}) = 0$ and restricting U and V to have unit variances, i.e

$$E(U^2) = 1 \Rightarrow \mathbf{a}'E(\mathcal{Y}'\mathcal{Y})\mathbf{a} = \mathbf{a}'\Omega_{yy} \otimes \Sigma_{yy}\mathbf{a} = 1 \tag{2.2.2}$$

$$E(V^2) = 1 \Rightarrow \mathbf{b}'E(\mathcal{X}'\mathcal{X})\mathbf{b} = \mathbf{b}'\Omega_{xx} \otimes \Sigma_{xx}\mathbf{b} = 1, \tag{2.2.3}$$

the correlation between U and V is given by

$$E(UV) = E(\mathbf{a}'\mathcal{Y}\mathcal{X}'\mathbf{b}) = \mathbf{a}'E(\mathcal{Y}\mathcal{X}')\mathbf{b} = \mathbf{a}'\Omega_{yx} \otimes \Sigma_{yx}\mathbf{b}. \tag{2.2.4}$$

Thus the algebraic problem is to find $\mathbf{a}$ and $\mathbf{b}$ to maximize 2.2.4 subject to the conditions 2.2.2 and 2.2.3.
Let

$$\psi = \mathbf{a}'\Omega_{yx} \otimes \Sigma_{yx}\mathbf{b} - \frac{\lambda}{2}(\mathbf{a}'\Omega_{yy} \otimes \Sigma_{yy}\mathbf{a} - 1) - \frac{\mu}{2}(\mathbf{b}'\Omega_{xx} \otimes \Sigma_{xx}\mathbf{b} - 1), \tag{2.2.5}$$

where λ and μ are Lagrange multipliers. Setting the partial derivatives of ψ with

respect to $\mathbf{a}$ and $\mathbf{b}$, equal to zero yields

$$\Omega_{yx} \otimes \Sigma_{yx}\mathbf{b} - \lambda\Omega_{yy} \otimes \Sigma_{yy}\mathbf{a} = 0, \tag{2.2.6}$$

$$\text{and } \Omega_{xy} \otimes \Sigma_{xy}\mathbf{a} - \mu\Omega_{xx} \otimes \Sigma_{xx}\mathbf{b} = 0. \tag{2.2.7}$$

Pre multiplication of equation 2.2.6 by $\mathbf{a}'$ and equation 2.2.7 by $\mathbf{b}'$ gives

$$\mathbf{a}'\Omega_{yx} \otimes \Sigma_{yx}\mathbf{b} - \lambda\mathbf{a}'\Omega_{yy} \otimes \Sigma_{yy}\mathbf{a} = 0, \tag{2.2.8}$$

$$\text{and } \mathbf{b}'\Omega_{xy} \otimes \Sigma_{xy}\mathbf{a} - \mu\mathbf{b}'\Omega_{xx} \otimes \Sigma_{xx}\mathbf{b} = 0. \tag{2.2.9}$$

Since $\mathbf{a}'\Omega_{yy} \otimes \Sigma_{yy}\mathbf{a} = 1$ and $\mathbf{b}'\Omega_{xx} \otimes \Sigma_{xx}\mathbf{b} = 1$ we get

$$\mathbf{a}'\Omega_{yx} \otimes \Sigma_{yx}\mathbf{b} - \lambda = 0, \tag{2.2.10}$$

$$\text{and } \mathbf{b}'\Omega_{xy} \otimes \Sigma_{xy}\mathbf{a} - \mu = 0. \tag{2.2.11}$$

This shows that

$$\lambda = \mu = \mathbf{a}'\Omega_{yx} \otimes \Sigma_{yx}\mathbf{b}.$$

Hence equations 2.2.6 and 2.2.7 can be written as

$$-\lambda\Omega_{yy} \otimes \Sigma_{yy}\mathbf{a} + \Omega_{yx} \otimes \Sigma_{yx}\mathbf{b} = 0, \tag{2.2.12}$$

$$\Omega_{xy} \otimes \Sigma_{xy}\mathbf{a} - \lambda\Omega_{xx} \otimes \Sigma_{xx}\mathbf{b} = 0. \tag{2.2.13}$$

Multiplying equation 2.2.12 by λ and premultiplying equation 2.2.13 by $(\Omega_{xx} \otimes \Sigma_{xx})^{-1}$ we get

$$-\lambda^2\Omega_{yy} \otimes \Sigma_{yy}\mathbf{a} + \Omega_{yx} \otimes \Sigma_{yx}\lambda\mathbf{b} = 0, \tag{2.2.14}$$

$$\text{and } (\Omega_{xx} \otimes \Sigma_{xx})^{-1}\Omega_{xy} \otimes \Sigma_{xy}\mathbf{a} = \lambda\mathbf{b}. \tag{2.2.15}$$

Combining these equations we get

$$-\lambda^2\Omega_{yy} \otimes \Sigma_{yy}\mathbf{a} + (\Omega_{yx} \otimes \Sigma_{yx})(\Omega_{xx} \otimes \Sigma_{xx})^{-1}(\Omega_{xy} \otimes \Sigma_{xy})\mathbf{a} = 0, \tag{2.2.16}$$

$$\text{i.e } \Big((\Omega_{yx} \otimes \Sigma_{yx})(\Omega_{xx} \otimes \Sigma_{xx})^{-1}(\Omega_{xy} \otimes \Sigma_{xy}) - \lambda^2\Omega_{yy} \otimes \Sigma_{yy}\Big)\mathbf{a} = 0. \tag{2.2.17}$$

It is clear from equation 2.2.17 that λ^2 is an eigenvalue of

$$\begin{aligned} A &= (\Omega_{yy} \otimes \Sigma_{yy})^{-1/2}(\Omega_{yx} \otimes \Sigma_{yx})(\Omega_{xx} \otimes \Sigma_{xx})^{-1}(\Omega_{xy} \otimes \Sigma_{xy})(\Omega_{yy} \otimes \Sigma_{yy})^{-1/2} \\ &= (\Omega_{yy}{}^{-1/2} \otimes \Sigma_{yy}{}^{-1/2})(\Omega_{yx} \otimes \Sigma_{yx})(\Omega_{xx}{}^{-1} \otimes \Sigma_{xx}{}^{-1})(\Omega_{xy} \otimes \Sigma_{xy})(\Omega_{yy}{}^{-1/2} \otimes \Sigma_{yy}{}^{-1/2}) \\ &= (\Omega_{yy}{}^{-1/2}\Omega_{yx}\Omega_{xx}{}^{-1}\Omega_{xy}\Omega_{yy}{}^{-1/2}) \otimes (\Sigma_{yy}{}^{-1/2}\Sigma_{yx}\Sigma_{xx}{}^{-1}\Sigma_{xy}\Sigma_{yy}{}^{-1/2}). \end{aligned}$$

Similarly we can show that equations 2.2.6 and 2.2.7 can be written as

$$\Omega_{yx} \otimes \Sigma_{yx}\mathbf{b} - \mu\Omega_{yy} \otimes \Sigma_{yy}\mathbf{a} = 0, \tag{2.2.18}$$

$$-\mu\Omega_{xx} \otimes \Sigma_{xx}\mathbf{b} + \Omega_{xy} \otimes \Sigma_{xy}\mathbf{a} = 0. \tag{2.2.19}$$

Multiplying equation 2.2.19 by μ and equation 2.2.18 by $(\Omega_{yy} \otimes \Sigma_{yy})^{-1}$ we get

$$(\Omega_{yy} \otimes \Sigma_{yy})^{-1}(\Omega_{yx} \otimes \Sigma_{yx})\mathbf{b} = \mu\mathbf{a}, \tag{2.2.20}$$

$$-\mu^2\Omega_{xx} \otimes \Sigma_{xx}\mathbf{b} + \mu\Omega_{xy} \otimes \Sigma_{xy}\mathbf{a} = 0. \tag{2.2.21}$$

As before combining these we get

$$(\Omega_{xy} \otimes \Sigma_{xy})(\Omega_{yy} \otimes \Sigma_{yy})^{-1}(\Omega_{yx} \otimes \Sigma_{yx})\mathbf{b} - \mu^2\Omega_{xx} \otimes \Sigma_{xx}\mathbf{b} = 0, \tag{2.2.22}$$

$$\Big((\Omega_{xy} \otimes \Sigma_{xy})(\Omega_{yy} \otimes \Sigma_{yy})^{-1}(\Omega_{yx} \otimes \Sigma_{yx}) - \mu^2\Omega_{xx} \otimes \Sigma_{xx}\Big)\mathbf{b} = 0, \tag{2.2.23}$$

where μ^2 is an eigenvalue of

$$\begin{aligned} B &= (\Omega_{xx} \otimes \Sigma_{xx})^{-1/2}(\Omega_{xy} \otimes \Sigma_{xy})(\Omega_{yy} \otimes \Sigma_{yy})^{-1}(\Omega_{yx} \otimes \Sigma_{yx})(\Omega_{xx} \otimes \Sigma_{xx})^{-1/2} \\ &= ({\Omega_{xx}}^{-1/2}\Omega_{xy}{\Omega_{yy}}^{-1}\Omega_{yx}{\Omega_{xx}}^{-1/2}) \otimes (\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1/2}). \end{aligned}$$

In general the vectors $\mathbf{a}_i$ and $\mathbf{b}_i$, such that $(\mathbf{a}_i'\mathcal{Y}, \mathbf{b}_i'\mathcal{X})$ is the i^{th} pair of canonical variables, are obtained as the solutions of

$$({\Omega_{yy}}^{-1/2}\Omega_{yx}{\Omega_{xx}}^{-1}\Omega_{xy}{\Omega_{yy}}^{-1/2}) \otimes ({\Sigma_{yy}}^{-1/2}\Sigma_{yx}{\Sigma_{xx}}^{-1}\Sigma_{xy}{\Sigma_{yy}}^{-1/2})\mathbf{a}_i = \lambda_i^2\mathbf{a}_i$$

and

$$({\Omega_{xx}}^{-1/2}\Omega_{xy}{\Omega_{yy}}^{-1}\Omega_{yx}{\Omega_{xx}}^{-1/2}) \otimes (\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1/2})\mathbf{b}_i = \lambda_i^2\mathbf{b}_i,$$

It is interesting to note that after fitting the repeated effect the canonical correlations are scaled by the eigenvalues of repeated effect matrix $({\Omega_{yy}}^{-1/2}\Omega_{yx}{\Omega_{xx}}^{-1}\Omega_{xy}{\Omega_{yy}}^{-1/2})$. It is possible to write $\boldsymbol{\lambda^2} = \boldsymbol{\lambda}_\Omega^2 \otimes \boldsymbol{\lambda}_\Sigma^2$, where $\boldsymbol{\lambda}_\Omega^2$ and $\boldsymbol{\lambda}_\Sigma^2$ are the eigenvalues of $({\Omega_{xx}}^{-1/2}\Omega_{xy}{\Omega_{yy}}^{-1}\Omega_{yx}{\Omega_{xx}}^{-1/2})$ and $({\Sigma_{yy}}^{-1/2}\Sigma_{yx}{\Sigma_{xx}}^{-1}\Sigma_{xy}{\Sigma_{yy}}^{-1/2})$ respectively.
Further, the vector $\mathbf{a}_i$ can be constructed by $\mathbf{a}_i = \mathbf{a}_i^\Omega \otimes \mathbf{a}_i^\Sigma$, where $\mathbf{a}_i^\Omega$ and $\mathbf{a}_i^\Sigma$ are the i^{th} eigenvectors of $({\Omega_{yy}}^{-1/2}\Omega_{yx}{\Omega_{xx}}^{-1}\Omega_{xy}{\Omega_{yy}}^{-1/2})$ and $({\Sigma_{yy}}^{-1/2}\Sigma_{yx}{\Sigma_{xx}}^{-1}\Sigma_{xy}{\Sigma_{yy}}^{-1/2})$ respectively. Similarly we can construct $\mathbf{b}_i$ from the corresponding matrices.
Also notice that if there is no repeated effect (that is, $\Omega_{ij} = I$, for $i, j = x, y$) or all the repeated effect is same (that is, $\Omega_{ij} = \Omega$) then

$$({\Omega_{xx}}^{-1/2}\Omega_{xy}{\Omega_{yy}}^{-1}\Omega_{yx}{\Omega_{xx}}^{-1/2}) = \omega\mathbf{I}_{tt}$$

and $\boldsymbol{\lambda}_\Omega^2 = \omega\mathbf{1_t}$, where ω is a positive constant.

II.3 Sample canonical correlations

Usually the matrices Σ_{yy}, Σ_{yx}, Σ_{xx}, Ω_{yy}, Ω_{yx} and Ω_{xx} are not known and need to be estimated from the data. The population canonical correlation will be estimated by the sample canonical correlations. Let us assume that $\mathbf{u} = (\mathcal{Y}', \mathcal{X}')'$ is distributed as multivariate normal with mean vector μ and variance covariance matrix $\mathbf{D}$.

Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be the random sample from the $N(\mu, \mathbf{D})$ where variance-covariance matrix $\mathbf{D}$ is given by equation 2.2.1.

The log-likelihood function of the parameters given the observed data is

$$L(\mu, \mathbf{D}) = -0.5(n\ log(|\mathbf{D}|) + \sum_{i=1}^{n}(\mathbf{u}_i - \mu)'\mathbf{D}^{-1}(\mathbf{u}_i - \mu)). \tag{2.3.1}$$

The estimates $\hat{\mu}$ and $\hat{\mathbf{D}}$ can be obtained by maximizing the above log-likelihood function. We used SAS non linear optimization routine for maximizing the log-likelihood function. Suppose $\hat{\Omega}_{yy}$, $\hat{\Omega}_{yx}$, and $\hat{\Omega}_{xx}$ are the maximum likelihood estimates of Ω_{yy}, Ω_{yx}, and Ω_{xx} respectively and $\hat{\Sigma}_{yy}$, $\hat{\Sigma}_{yx}$, and $\hat{\Sigma}_{xx}$ are the maximum likelihood estimates of Σ_{yy}, Σ_{yx}, and Σ_{xx} respectively.

Then the sample canonical correlations $r_1 \geq r_2 \geq \ldots \geq r_p$ are obtained as the positive square roots of the nonzero eigenvalues of

$$(\hat{\Omega}_{yy}^{-1/2}\hat{\Omega}_{yx}\hat{\Omega}_{xx}^{-1}\hat{\Omega}_{xy}\hat{\Omega}_{yy}^{-1/2}) \otimes (\hat{\Sigma}_{yy}^{-1/2}\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1/2}).$$

The vectors $\hat{\mathbf{a}}_i$ and $\hat{\mathbf{b}}_i$ corresponding to i^{th} pair of canonical variables are obtained as the solution of $(\hat{\Omega}_{yy}^{-1/2}\hat{\Omega}_{yx}\hat{\Omega}_{xx}^{-1}\hat{\Omega}_{xy}\hat{\Omega}_{yy}^{-1/2}) \otimes (\hat{\Sigma}_{yy}^{-1/2}\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1/2})\hat{\mathbf{a}}_i = \hat{\lambda}_i^2\hat{\mathbf{a}}_i$ and

$$(\hat{\Omega}_{xx}^{-1/2}\hat{\Omega}_{xy}\hat{\Omega}_{yy}^{-1}\hat{\Omega}_{yx}\hat{\Omega}_{xx}^{-1/2}) \otimes (\hat{\Sigma}_{xx}^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1}\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1/2})\hat{\mathbf{b}}_i = \hat{\lambda}_i^2\hat{\mathbf{b}}_i.$$

II.4 Hypothesis

Before performing any canonical correlation analysis using the samples $\mathbf{u}_1, \ldots, \mathbf{u}_n$, the following hypotheses may be tested.

1. First test for the repeated effect on the variance covariance matrices of $\mathbf{y}$, of $\mathbf{x}$ and on $cov(\mathbf{x}, \mathbf{y})$, i.e. test

$$H_0 : D = \begin{bmatrix} I_{yy} \otimes \Sigma_{yy} & I_{yx} \otimes \Sigma_{yx} \\ I_{xy} \otimes \Sigma_{xy} & I_{xx} \otimes \Sigma_{xx} \end{bmatrix} \text{ vs } H_a : D = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix}$$

Note that the null hypothesis here specifies that the variance and covariance matrices do not change with the time factor. Here as well as in the cases that follow, the alternative hypothesis is assumed to be as in our assumed model, that it is unstructured Kronecker product block matrix. Testing can be performed using the likelihood ratio test (LRT) statistic. Maximizing the log-likelihood function $L(\mu, \mathbf{D}) = -0.5(n\ log(|\mathbf{D}|) + \sum_{i=1}^{n}(\mathbf{u}_i - \mu)'\mathbf{D}^{-1}(\mathbf{u}_i - \mu))$ under H_0 and H_a will produce the maximum likelihood estimates. The likelihood ratio test statistic is then

$$-2log\Lambda = -2log(\ell_0/\ell_a),$$

where ℓ_0 and ℓ_a denote the maximized likelihood functions under the null and alternative hypothesis. Under H_0, $-2log\Lambda$ has a chi-squared distribution, as $n \to \infty$. The degrees of freedoms equal to the difference in the dimensions of the parameter spaces under $H_0 \cup H_a$ and under H_0.

2. If we accept H_0 then we can do the usual canonical correlation analysis by merging all the data. Otherwise we will test whether the effect of time (or the repeated effect) is on the covariances between ($\mathbf{x}$ and $\mathbf{y}$) only. This amounts to testing

$$H_{01} : D = \begin{bmatrix} I_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & I_{xx} \otimes \Sigma_{xx} \end{bmatrix} \text{ vs } H_a : D = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix}$$

To test this hypothesis, ℓ_a is as in the previous case, i.e. as in (1) above. The maximum likelihood estimate and the maximum value of the likelihood under H_{01} can be obtained by maximizing 2.3.1 under H_{01}.

3. If we accept H_{01}, then we can perform canonical correlation analysis (CCA) using the estimated variance covariance matrix given under H_{01} in (2). above. Otherwise we will test for repeated effect on variance covariance matrices of $\mathbf{y}$, $\mathbf{x}$, by testing,

$$H_{ox} : D = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & I_{xx} \otimes \Sigma_{xx} \end{bmatrix} \text{ vs } H_a : D = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix},$$

$$H_{oy} : D = \begin{bmatrix} I_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix} \text{ vs } H_a : D = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix}$$

As before the MLE and the value of the maximum likelihood function under H_{ox} (and H_{oy}) can be obtained by maximizing 2.3.1 under the null hypothesis. Under H_a, the value ℓ_a remains the same.

4. If we accept H_{ox} or H_{oy} then we can perform canonical correlation analysis (CCA) using the corresponding estimated variance covariance matrix as in (3). Otherwise we will test for the same repeated effect, that is, test

$$H_{tt} : D = \begin{bmatrix} \Omega_{tt} \otimes \Sigma_{yy} & \Omega_{tt} \otimes \Sigma_{yx} \\ \Omega_{tt} \otimes \Sigma_{xy} & \Omega_{tt} \otimes \Sigma_{xx} \end{bmatrix} \text{ vs } H_a : D = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix}$$

The MLE of the common Ω_{tt} and the other parameters can be obtained by maximizing 2.3.1 under H_{tt} and in the same way as before the LRT can be constructed.

5. If we accept H_{tt} then it suggest that change in variance covariance matrices over time is same and we should perform canonical correlation analysis (CCA) using the estimated structured variance covariance matrix as discussed in (4) above. Otherwise we should proceed with the general structured variance covariance matrix

$$\mathbf{D} = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix}.$$

II.5 Constructing a Variance Covariance Matrix for Simulation

In order to illustrate the analysis discussed here, we will work with simulated data. First we use the Helmert matrix to generate the positive definite matrices. The general form of a Helmert matrix $\mathbf{H}_k$ of order k has $k^{-1/2}\mathbf{1}_k'$ for its first row, and each of its other $k-1$ rows for $i = 1, \ldots, k-1$ has the partitioned form

$$\left[\begin{array}{c|c|c} \mathbf{1}_i' & -i & 0 \end{array} \right] / \sqrt{\lambda_i}$$

with $\lambda_i = i(i+1)$. A Helmert matrix is an orthogonal matrix, that is, $\mathbf{H'H} = \mathbf{HH'} = \mathbf{I}_k$. For example, the 4^{th} order Helmert matrix is given by

$$\mathbf{H}_4 = \begin{bmatrix} \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{-3}{\sqrt{12}} \end{bmatrix}.$$

The spectral decomposition of a symmetric matrix, $\mathbf{A}$ is $\mathbf{A} = \sum \lambda_i \mathbf{u}_i \mathbf{u}_i'$, where the $\mathbf{u}_i$'s are the eigenvectors of $\mathbf{A}$.

Now to generate a $k \times k$ positive definite matrix we take the k^{th} order Helmert matrix, whose columns will give us the eigenvector of the desired matrix. Then choosing k positive eigenvalues and using the spectral decomposition property we can construct the desired $k \times k$ positive definite matrix. We will use thus constructed positive definite matrix as $\mathbf{\Sigma}$. Partitioning $\mathbf{\Sigma}$ will give

$$\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix},$$

and Σ_{yy}, Σ_{xx} and Σ_{yx} can be used as variance-covariance matrix for $\mathbf{y}$, $\mathbf{x}$ and covariance matrix between $\mathbf{y}$ and $\mathbf{x}$ respectively. Then by choosing $t \times t$ modeling matrix Ω_{yy} to associate with Σ_{yy}, Ω_{xx} with Σ_{xx} and Ω_{yx} with Σ_{yx} we can construct the desired matrix

$$\mathbf{D} = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix}$$

We can simulate any desired number of observations from the multivariate Normal $N(\mathbf{0}, \mathbf{D})$ and do repeated canonical correlation analysis on them as discussed in section II.3.

II.6 Results and Discussion

To conduct a simulation purpose we chose three $\mathbf{y}$ components, two $\mathbf{x}$ components and three repeated measurements on those i.e. $q = 3$, $p = 2$, and $t = 3$. A Helmert matrix of order 5 is chosen and used to determine a 5×5 positive definite variance covariance matrix Σ. In addition the following eigenvalues are arbitrarily selected,

9.5262261, 8.7983733, 4.3901993, 2.2263795 and 1.9919697 to generate a positive definite matrix Σ by the method described earlier

$$\Sigma = \begin{bmatrix} 5.3866296 & 2.1523738 & 0.3669639 & -1.729692 & 1.580125 \\ 2.1523738 & 5.3951715 & 1.9648395 & 1.3893513 & 1.0761869 \\ 0.3669639 & 1.9648395 & 4.7251901 & 0.2368742 & 0.183482 \\ -1.729692 & 1.3893513 & 0.2368742 & 8.4097148 & -0.864846 \\ 1.580125 & 1.0761869 & 0.183482 & -0.864846 & 3.016442 \end{bmatrix}.$$

By partitioning Σ we get Σ_{yy}, Σ_{xx} and Σ_{yx} as follows:

$$\Sigma_{yy} = \begin{bmatrix} 5.3866296 & 2.1523738 & 0.3669639 \\ 2.1523738 & 5.3951715 & 1.9648395 \\ 0.3669639 & 1.9648395 & 4.7251901 \end{bmatrix},$$

$$\Sigma_{xx} = \begin{bmatrix} 8.4097148 & -0.864846 \\ -0.864846 & 3.016442 \end{bmatrix} \text{ and } \Sigma_{yx} = \begin{bmatrix} -1.729692 & 1.580125 \\ 1.3893513 & 1.0761869 \\ 0.2368742 & 0.183482 \end{bmatrix}.$$

We assume $AR(1)$ structure for repeated modeling matrices Ω_{yy}, Ω_{xx}, and Ω_{yx} with correlation parameter $\rho_y = 0.1$, $\rho_x = 0.2$, and $\rho_{yx} = 0.1$ respectively. Arranging all the matrices together we have

$$\mathbf{D} = \begin{bmatrix} \Omega_{yy} \otimes \Sigma_{yy} & \Omega_{yx} \otimes \Sigma_{yx} \\ \Omega_{xy} \otimes \Sigma_{xy} & \Omega_{xx} \otimes \Sigma_{xx} \end{bmatrix}.$$

We simulated 500 observations from the multivariate Normal $N(\mathbf{0}, \mathbf{D})$ and estimated the population parameters $\hat{\Sigma}_{yy}$, $\hat{\Sigma}_{xx}$, $\hat{\Sigma}_{yx}$, $\hat{\rho}_y$, $\hat{\rho}_x$, and $\hat{\rho}_{yx}$. The estimates were found by maximizing the log-likelihood function using SAS *NLPQN* optimization routine.

To illustrate the idea of hypothesis testing we used a data set generated from one of the simulations and tabulated the chi square test statistics values. P-values for testing different hypothesis are shown in Table 2.1. As can be seen from the Table 2.1, all of the p-values are quite small except for the H_{tt} hypothesis ($p-val = 0.1121743$). Thus all hypotheses except the H_{tt} are rejected. In hypothesis H_{tt} we are testing that the repeated effect is same on all components. In our simulation we have used the $AR(1)$ structure for the repeated correlation matrix with correlation parameter $\rho_y = 0.1$, $\rho_x = 0.2$, and $\rho_{yx} = 0.1$. Apparently these values are not very different to

reject H_{tt} using likelihood ratio test and this sample data. However, when we chose quite different AR(1) parameters, the LRT did reject H_{tt}.

We repeated the above procedure 5000 times and calculated the average values of estimates. Table 2.3 shows the average of the parameter estimates based on these simulations. Table 2.2 presents the mean of sample canonical correlations' estimates. At the left of estimates we have provided true parameter values. In Table 2.2, minimum and maximum bias values are 0.001382172 and 0.011013628 respectively. Similarly in Table 2.3 biases ranges from $4.23009E-05$ to 0.00409767. From both the tables it can be said that the estimates are very close to the true values.

II.7 Concluding Remarks

In this chapter, we have provided an easy to implement procedure to perform canonical correlation analysis of repeatedly observed data sets. To accommodate the effects of repeated measure we have adopted a Kronecker product structure to the variance covariance matrices. To account for the existence of repeated measure effects on different blocks of the variance covariance matrix, we have provided testing of different hypothesis. All of the procedures have been implemented on simulated data sets.

Table 2.1: Hypothesis Testing

Hypothesis	Chi Square Test Statistics	Dof	p-value
H_0	108.0483	3	0
H_{01}	85.490536	2	0
H_{ox}	53.496484	1	2.59E-13
H_{oy}	44.505918	1	2.54E-11
H_{tt}	4.3754035	2	0.1121743

Table 2.2: General Structure Correlation Estimates

Can. Corr.	Parameter ρ	Estimate $\hat{\rho}$	Root MSE $\sqrt{E((\rho-\hat{\rho})^2)}$	Bias $\lvert(\rho-\hat{\rho})\rvert$
ρ_1	0.237273404	0.2418999	0.025882426	0.004626013
ρ_2	0.208651984	0.2124004	0.018033303	0.003754997
ρ_3	0.177922968	0.1889365	0.020921281	0.011013628
ρ_4	0.1591106	0.159546475	0.018627936	0.009110434
ρ_5	0.14793029	0.1447519	0.016281892	0.00317805
ρ_6	0.126144001	0.1247618	0.018398369	0.001382172

Table 2.3: General Structure Estimates

Pop. Para.	**Para.** θ	**Estimate** $\hat{\theta}$	**Root MSE** $\sqrt{E((\theta-\hat{\theta})^2)}$	**Bias** $\lvert(\theta-\hat{\theta})\rvert$
$\Sigma_{yy}(1,1)$	5.38663	5.386232775	0.196693162	0.000397225
$\Sigma_{yy}(2,2)$	5.39517	5.394642743	0.201315673	0.000527257
$\Sigma_{yy}(3,3)$	4.72519	4.724343576	0.175820078	0.000846424
$\Sigma_{yy}(1,2)$	2.15237	2.149807789	0.151351577	0.00256221
$\Sigma_{yy}(1,3)$	0.36696	0.367644263	0.131625226	0.000684262
$\Sigma_{yy}(2,3)$	1.96484	1.963606497	0.139320494	0.001233503
$\Sigma_{xx}(1,1)$	8.40971	8.406462064	0.312531758	0.003247938
$\Sigma_{xx}(2,2)$	3.01644	3.014373036	0.111293755	0.002066964
$\Sigma_{xx}1,2$	-0.86485	-0.864892301	0.131922326	4.23009E-05
$\Sigma_{yx}(1,1)$	-1.72969	-1.727133373	0.177219073	0.002556627
$\Sigma_{yx}(1,2)$	1.58013	1.57732949	0.111941503	0.002800511
$\Sigma_{yx}(2,1)$	1.38935	1.387468714	0.175657622	0.001881287
$\Sigma_{yx}(2,2)$	1.07619	1.073755928	0.109225913	0.002434073
$\Sigma_{yx}(3,1)$	0.23687	0.232772331	0.159496395	0.00409767
$\Sigma_{yx}(3,2)$	0.18348	0.182754843	0.097614036	0.000725157
ρ_y	0.1	0.100341533	0.024503061	0.000341533
ρ_x	0.2	0.199774535	0.034666987	0.000225466
ρ_{yx}	0.1	0.100673171	0.041535527	0.000673171

CHAPTER III

REPEATED CORRESPONDENCE ANALYSIS

III.1 Introduction

Correspondence analysis (CA) is a graphical multivariate technique for performing an exploratory data analysis of a contingency table. The main problem of interest in CA is that of graphically representing rows and columns of a contingency table as points in a lower dimensional Euclidean space such that the affinities of the rows or columns in the higher dimensional space are preserved as much as possible in the lower dimensional Euclidean space. The graph is then used to gain understanding of the data and to extract information from it. The graphs in correspondence analysis can be used to determine, to some extent, the possible association between the two sets of variable. CA is used frequently to determine those categories of a variable that are similar.

In the next section, for the benefit of introducing the notation, we will briefly review canonical correspondence analysis. More details about the method and its applications can be found in many books. For example, see Greenacre (1984) and Khattree and Naik (2000). However, the main focus in this chapter is to extend correspondence analysis to repeated measures data. First we illustrate how correspondence analysis can be viewed as CCA of the previous chapter and then we will provide methods for performing an analysis of longitudinally observed contingency tables.

III.2 Correspondence Analysis

Let $\mathbf{X}$ and $\mathbf{Y}$ denote two categorical variables with a and b categories respectively. Let $\mathbf{N}$ be $a \times b$ contingency table with frequency, $n_{ij} \geq 0$ in the $(i,j)^{th}$ cell. The correspondence matrix $\mathbf{P}$ is defined as the matrix of elements of $\mathbf{N}$ divided by the grand total , that is,

$$\mathbf{P}_{a\times b} = (p_{ij}) = (\frac{n_{ij}}{n}), \text{ where } n = \sum_i \sum_j n_{ij}.$$

The correspondence matrix along with the row and column marginal totals can be displayed as

p_{11}	p_{12}	. . .	p_{1b}	$p_{1.}$
p_{21}	p_{22}	. . .	p_{2b}	$p_{2.}$
.	.		.	.
.	.	. . .	.	.
.	.		.	.
p_{a1}	p_{a2}	. . .	p_{ab}	$p_{a.}$
$p_{.1}$	$p_{.2}$	. . .	$p_{.b}$	1

Let the vector of row sums of $\mathbf{P}$ be $\mathbf{r} = \mathbf{P1} = (p_{1.}, \ldots, p_{a.})'$ and the vector of column sums of $\mathbf{P}$ be $\mathbf{c} = \mathbf{P'1} = (p_{.1}, \ldots, p_{.b})'$. Let

$$\mathbf{D}_r = diag(\mathbf{r}) == \begin{bmatrix} p_{1.} & 0 & \ldots & 0 \\ 0 & p_{2.} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_{a.} \end{bmatrix}$$

and

$$\mathbf{D}_c = diag(\mathbf{c}) == \begin{bmatrix} p_{.1} & 0 & \ldots & 0 \\ 0 & p_{.2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_{.b} \end{bmatrix}.$$

Then the row-profiles in the b-dimensional space are given by

$$\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P} \equiv \begin{bmatrix} \tilde{\mathbf{r}}_1' \\ \vdots \\ \tilde{\mathbf{r}}_a' \end{bmatrix}$$

and column-profiles in the a-dimensional space are given by

$$\mathbf{C} = \mathbf{D}_c^{-1}\mathbf{P}' \equiv \begin{bmatrix} \tilde{\mathbf{c}}_1' \\ \vdots \\ \tilde{\mathbf{c}}_b' \end{bmatrix}.$$

The row and column profiles define two clouds of points in respective b- and a-dimensional Euclidean spaces. The centroids of row and column clouds in their respective spaces are

$$\text{Row centroid: } \mathbf{c} = \mathbf{R}'\mathbf{r} \qquad \text{Column centroid: } \mathbf{r} = \mathbf{C}'\mathbf{c}.$$

Note: $\mathbf{R}'\mathbf{r} = \mathbf{P}'\mathbf{D}_r^{-1}\mathbf{r} = \mathbf{P}'\mathbf{1} = \mathbf{c}$ and $\mathbf{C}'\mathbf{c} = \mathbf{P}\mathbf{D}_c^{-1}\mathbf{c} = \mathbf{P}\mathbf{1} = \mathbf{r}$.

The overall spatial variation of each cloud of points is quantified by their total inertia, that is, the weighted sum of squared distances from the points to their respective centroids,

$$in(a) = \sum_i r_i(\tilde{\mathbf{r}}_i - \mathbf{c})'\mathbf{D}_c^{-1}(\tilde{\mathbf{r}}_i - \mathbf{c})$$

and

$$in(b) = \sum_i c_i(\tilde{\mathbf{c}}_i - \mathbf{r})'\mathbf{D}_r^{-1}(\tilde{\mathbf{c}}_i - \mathbf{r}),$$

where $in(a)$ and $in(b)$ are total inertia of row profiles and column profiles respectively. Also, c_i and r_i are the i^{th} elements of the vectors $\mathbf{c}$ and $\mathbf{r}$ respectively. Both clouds have the same total inertia and n times it is equal to the chi-square statistic for "independence," that is,

$$in(a) = in(b) \equiv trace[\mathbf{D}_r^{-1}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1}(\mathbf{P} - \mathbf{r}\mathbf{c}')'] = \chi^2/n.$$

A lower dimensional space, say the k^* -dimensional subspace, of the row and column clouds which are closest to the points in terms of weighted sum of squared distances are determined using generalized singular value decomposition of the matrix $(\mathbf{P} - \mathbf{r}\mathbf{c}')$, that is given by

$$(\mathbf{P} - \mathbf{r}\mathbf{c}') = \mathbf{A}\mathbf{\Lambda}\mathbf{B}', \tag{3.2.1}$$

where matrix $\mathbf{A}_{a\times m}$ and $\mathbf{B}_{b\times m}$ are such that $\mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A} = \mathbf{I}_m$ and $\mathbf{B}'\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}_m$ and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the singular values $\lambda_1, \ldots, \lambda_m$ of $(\mathbf{P} - \mathbf{r}\mathbf{c}')$. The matrices $\mathbf{A}$ and $\mathbf{B}$ can be obtained from the usual singular value decomposition of $\mathbf{T} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}$. Note that $\lambda_1^2, \ldots, \lambda_m^2$ are the eigenvalues of $\mathbf{T}\mathbf{T}'$.

In practice the value of k^* is taken to be 2 or 3. The coordinates for the a row profiles are the a rows of the matrix formed by taking the first k^* columns of

$$\mathbf{F} = \mathbf{D}_r^{-1}\mathbf{A}\mathbf{\Lambda} \tag{3.2.2}$$

and for b column profiles are the b rows of the matrix formed by taking the first k^* columns of

$$\mathbf{G} = \mathbf{D}_c^{-1}\mathbf{B}\mathbf{\Lambda}. \tag{3.2.3}$$

These coordinates are generally plotted on a plane on the same graph. This kind of display is called symmetric plot (Greenacre, 1984). In this plot the distance between the points corresponding to the row profiles or those points corresponding to column profiles are the approximation to the corresponding chi-square distances between the respective profiles. But the distance between two points, one corresponding to a row profile and another corresponding to a column profile, has no such interpretation. In the following example we use a real life data set to illustrate correspondence analysis.

III.2.1 An Example

The data considered here are from Srole, Langner, Michael, Kirkpatrick, Opler and Rennie (1978) and given in Table 3.1. The objective of the study is to examine the relationship, if any, between children's mental impairment and parent's socioeconomic status. There are six levels of socioeconomic status from 1 (high) to 6 (low) and four levels of mental health status: Well, mild symptom formation (MILD), moderate symptom formation (MODERATE) and impairment (IMPAIRED). Data obtained in the form of 6 by 4 contingency table are based on a sample of 1660 residents of Manhattan.

Correspondence analysis of these data is shown in Khattree and Naik (2000). For testing the null hypothesis of no association between the parent's socioeconomic status and children's mental impairment, chi-square test statistics resulted in 45.9853 with 15 degree of freedom. Chi-square decomposition is given in Table 3.2. The small P-value (< 0.00001) suggests that we reject the null hypothesis and conclude that parent's socioeconomic status and children's mental impairment are not independent. The correspondence matrix $\mathbf{P}$ for these data is given by

$$\mathbf{P} = \begin{bmatrix} 0.0385542 & 0.0566265 & 0.0349398 & 0.0277108 \\ 0.0343373 & 0.0566265 & 0.0325301 & 0.0240964 \\ 0.0343373 & 0.0632530 & 0.0391566 & 0.0361446 \\ 0.0433735 & 0.0849398 & 0.0463855 & 0.0566265 \\ 0.0216867 & 0.0584337 & 0.0325301 & 0.0469880 \\ 0.0126506 & 0.0427711 & 0.0325301 & 0.0427711 \end{bmatrix}$$

The co-ordinates of parent's socioeconomic status and children's mental impairment in two-dimensional space are given by $\mathbf{F}_{6\times2}$ and $\mathbf{G}_{4\times2}$ respectively, and they are

$$\mathbf{F}_{6\times2} = \begin{bmatrix} 0.1809 & 0.0192 \\ 0.1850 & 0.0116 \\ 0.0590 & 0.0222 \\ -0.0089 & -0.0421 \\ -0.1654 & -0.0436 \\ -0.2877 & 0.0620 \end{bmatrix} \qquad \mathbf{G}_{4\times2} = \begin{bmatrix} 0.2595 & -0.0121 \\ 0.0296 & -0.0237 \\ -0.0142 & 0.0699 \\ -0.2374 & -0.0189 \end{bmatrix}$$

Figure 3.1 is a two-dimensional plot generated by correspondence analysis of the socioeconomic status by mental health of children data. For the first dimension the value of the principal inertia is $\lambda_1^2 = 0.0260$. The percentage of total inertia explained by the one-dimensional approximation is approximately 94%. This percentage explained by two-dimensional approximation is close 99%. Since the whole space here is three-dimensional we can be confident that the two-dimensional representation of the row profiles will be a reasonably good approximation to the whole space. In this case, categories are ordered and the order is maintained along the first principal axis. Categories 1 and 2 cannot be clearly distinguished hence it may be clubbed together to form one group. The two middle categories corresponding to the mental status of children are quite close to each other, but there is a clear distinction between the other categories.

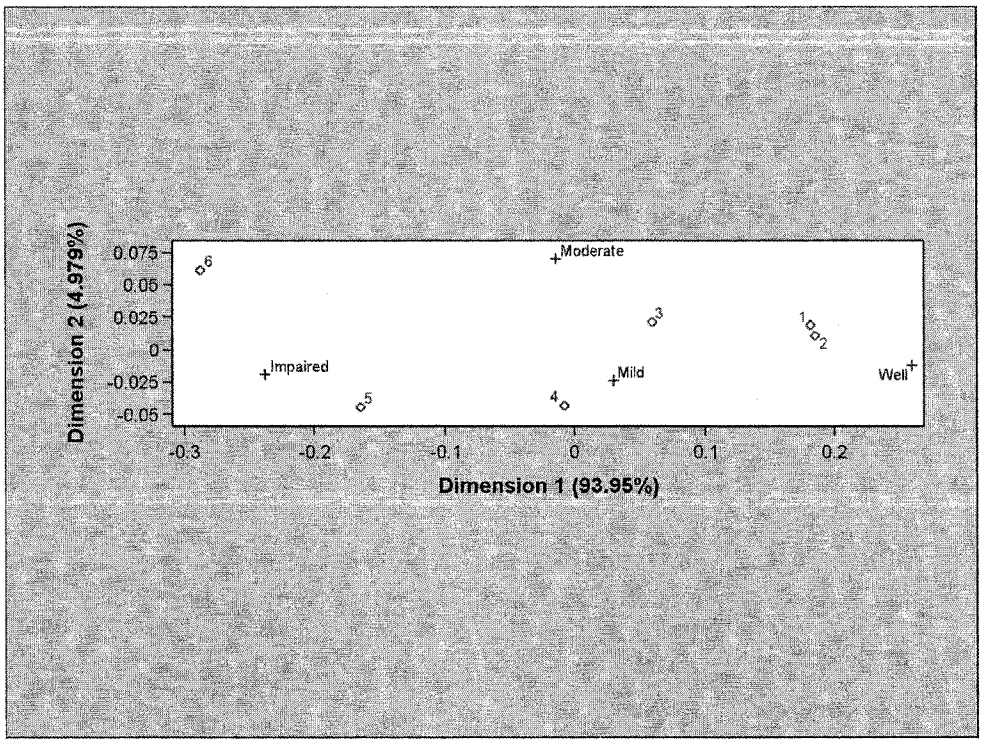

Figure 3.1: CA: Two-Dim Plot of Socioeconomic Status by Mental Health Data.

III.3 Correspondence Analysis as a Canonical Correlation Analysis

In the following, we illustrate how correspondence analysis can be formulated as canonical correlation analysis. See Goodman (1981) and O'Neill (1981).

Let $\mathbf{X} = (X_1, ..., X_a)'$ and $\mathbf{Y} = (Y_1, ..., Y_b)'$ be two categorical variables with a and b categories respectively. Consider

	Y_1	Y_2	...	Y_b	
X_1	p_{11}	p_{12}	$\cdots$	p_{1m}	$p_{1.}$
X_2	p_{21}	p_{22}	$\cdots$	p_{2m}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
X_a	p_{n1}	p_{n2}	$\cdots$	p_{nm}	$p_{a.}$
	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.b}$	1

where p_{ij} is the probability of $\mathbf{X}$ assuming the i^{th} category and $\mathbf{Y}$ assuming the j^{th} category, $p_{.j}$ is the marginal probability that $\mathbf{Y}$ assumes j^{th} category and $p_{i.}$ is the

marginal probability that $\mathbf{X}$ assumes the i^{th} category. Then the variance covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$D\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx}\\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx}\end{bmatrix}, \tag{3.3.1}$$

where

$$\boldsymbol{\Sigma}_{yy} = \begin{bmatrix} p_{.1}(1-p_{.1}) & -p_{.1}p_{.2} & \cdots & -p_{.1}p_{.b} \\ -p_{.1}p_{.2} & p_{.2}(1-p_{.2}) & \cdots & -p_{.2}p_{.b} \\ \vdots & \vdots & \vdots & \vdots \\ -p_{.1}p_{.b} & -p_{.2}p_{.b} & \cdots & p_{.b}(1-p_{.b}) \end{bmatrix} \tag{3.3.2}$$

$$\boldsymbol{\Sigma}_{xx} = \begin{bmatrix} p_{1.}(1-p_{1.}) & -p_{1.}p_{2.} & \cdots & -p_{1.}p_{a.} \\ -p_{1.}p_{2.} & p_{2.}(1-p_{2.}) & \cdots & -p_{2.}p_{a.} \\ \vdots & \vdots & \vdots & \vdots \\ -p_{1.}p_{a.} & -p_{2.}p_{a.} & \cdots & p_{a.}(1-p_{a.}) \end{bmatrix} \tag{3.3.3}$$

and

$$\boldsymbol{\Sigma}_{xy} = \begin{bmatrix} p_{11}-p_{1.}p_{.1} & p_{12}-p_{1.}p_{.2} & \cdots & p_{1b}-p_{1.}p_{.b} \\ p_{21}-p_{2.}p_{.1} & p_{22}-p_{2.}p_{.2} & \cdots & p_{2b}-p_{2.}p_{.b} \\ \vdots & \vdots & \vdots & \vdots \\ p_{a1}-p_{a.}p_{.1} & p_{a2}-p_{a.}p_{.2} & \cdots & p_{ab}-p_{a.}p_{.b} \end{bmatrix}. \tag{3.3.4}$$

Now performing canonical correlation analysis on this variance covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ will result in canonical variables of $\mathbf{X}$ and $\mathbf{Y}$, that are highly correlated. Canonical correlations are the square root of the eigenvalues of

$$\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-} \text{ or } \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-},$$

where $\boldsymbol{\Sigma}_{yy}^{-}$ and $\boldsymbol{\Sigma}_{xx}^{-}$ are the generalized inverses of $\boldsymbol{\Sigma}_{yy}$ and $\boldsymbol{\Sigma}_{xx}$, respectively. All of the population parameters are estimated by the corresponding sample counterparts. The first and second dimension coordinates of $\mathbf{X}$ and $\mathbf{Y}$ are the canonical coefficients of first and second dimension canonical variables.

The estimated variance covariance matrix of socioeconomic status by mental health of children data considered in section III.2.1 as given by equation 3.3.1 is

given by

$$\hat{\Sigma} = \left[\begin{array}{cccc|cccccc} 0.15 & -0.07 & -0.04 & -0.04 & 0.01 & 0.01 & 0.00 & 0.00 & -0.01 & -0.01 \\ -0.07 & 0.23 & -0.08 & -0.09 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.04 & -0.08 & 0.17 & -0.05 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.04 & -0.09 & -0.05 & 0.18 & -0.01 & -0.01 & 0.00 & 0.00 & 0.01 & 0.01 \\ \hline 0.01 & 0.00 & 0.00 & -0.01 & 0.13 & -0.02 & -0.03 & -0.04 & -0.03 & -0.02 \\ 0.01 & 0.00 & 0.00 & -0.01 & -0.02 & 0.13 & -0.03 & -0.03 & -0.02 & -0.02 \\ 0.00 & 0.00 & 0.00 & 0.00 & -0.03 & -0.03 & 0.14 & -0.04 & -0.03 & -0.02 \\ 0.00 & 0.00 & 0.00 & 0.00 & -0.04 & -0.03 & -0.04 & 0.18 & -0.04 & -0.03 \\ -0.01 & 0.00 & 0.00 & 0.01 & -0.03 & -0.02 & -0.03 & -0.04 & 0.13 & -0.02 \\ -0.01 & 0.00 & 0.00 & 0.01 & -0.02 & -0.02 & -0.02 & -0.03 & -0.02 & 0.11 \end{array}\right]. \quad (3.3.5)$$

Canonical correlations obtained from canonical correlation analysis is shown in Table 3.3.

The two dimensional coordinates for socioeconomic status and mental health status obtained from canonical correlation analysis approach and correspondence analysis (CA) approach are shown in Table 3.4 and 3.5 respectively. It is very clear from values in these tables that by performing canonical correlation analysis on variance covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ we will get results similar to that of correspondence analysis.

III.4 Repeated Correspondence Analysis

In this section we will show how to perform correspondence analysis if we have a repeated contingency table. Suppose, as before $\mathbf{X}$ and $\mathbf{Y}$ are two categorical variables with a and b categories and are observed over t time periods. Our data then constitute t contingency tables, $N_1, N_2, \ldots, N_t$, each is of $a \times b$ dimension, that is,

$$N_k = (y_{ijk}), \quad i = 1, \ldots, a; \; j = 1, \ldots, b; \; k = 1, \ldots, t$$

in which y_{ijk} denotes the frequency of the i^{th} category of $\mathbf{X}$ and j^{th} category of $\mathbf{Y}$ in the k^{th} contingency table.

Performing correspondence analysis (CA) on these tables and restricting the analysis to two dimensions, for each time period k ($k = 1, \ldots, t$) we get the quantities,

λ_{1k}, λ_{2k}, the first and second singular values; $\mathbf{D}_{1k}$, $\mathbf{D}_{2k}$, coordinates for plotting $a+b$ points on a two-dimensional plane; $\mathbf{U}_{1k}$, $\mathbf{V}_{1k}$, the principal axis for row; and $\mathbf{U}_{2k}$, $\mathbf{V}_{2k}$, the principal axis for column. The primary objective of the analysis is to see if there is any repeated measure effect on t correspondence analyses. If there is no repeated effect then we can merge t contingency tables into one and perform correspondence analysis (CA) of that table. Otherwise we have to interpret the result of correspondence analysis (CA) of each contingency table separately.

To assess the repeated effect we use the general linear modeling framework. We fit a general linear model with correlated errors to each of the above quantities. For example, a general linear model with correlated errors for the t first singular values is given by

$$\lambda_{1k} = \beta k + \epsilon_k \quad k = 1, \ldots, t, \tag{3.4.1}$$

where ϵ_k are the correlated random errors for $k = 1, \ldots, t$. Why such a model would be reasonable is clear from the results in O'Neill (1981). Since we have only one first singular value from each time period we have to assume a certain autoregressive type of structure to model the correlations among the errors.

Next, we discuss estimation of β and correlation parameter ρ for the model in 3.4.1, when $\boldsymbol{\epsilon} \sim N(0, \sigma^2 V(\rho))$, where $V(\rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{t-1} \\ \rho & 1 & \rho & \cdots & \rho^{t-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \rho^{t-3} & \cdots & 1 \end{bmatrix}$, an $AR(1)$ structure and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_t)'$.

The log likelihood function of the parameters, given $\boldsymbol{\lambda}_1 = (\lambda_1, ..., \lambda_t)'$, is given by

$$logf(\beta, \sigma^2, \rho | \boldsymbol{\lambda}_1) = \frac{t}{2} log(2\pi) - \frac{t-1}{2} log(1-\rho^2) - \frac{t}{2} log(\sigma^2) - \frac{1}{2\sigma^2} (\boldsymbol{\lambda}_1 - \beta \mathbf{T})' \Sigma^{-1} (\boldsymbol{\lambda}_1 - \beta \mathbf{T}), \tag{3.4.2}$$

where $\mathbf{T} = (1, \ldots, t)'$ and $\Sigma^{-1} = \frac{1}{(1-\rho^2)} \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1+\rho^2 & -\rho \\ 0 & \cdots & \cdots & -\rho & 1 \end{bmatrix}$.

Differentiating the log likelihood function with respect to β and σ^2 and equating to zero will give the following maximum likelihood estimating equations for β and σ^2

respectively.

$$\hat{\beta} = \frac{\mathbf{T}'\Sigma^{-1}\boldsymbol{\lambda}_1}{\mathbf{T}'\Sigma^{-1}\mathbf{T}} \tag{3.4.3}$$

$$\hat{\sigma^2} = \frac{(\boldsymbol{\lambda}_1 - \hat{\beta}\mathbf{T})'\Sigma^{-1}(\boldsymbol{\lambda}_1 - \hat{\beta}\mathbf{T})}{t}. \tag{3.4.4}$$

Let $\mathbf{e} = \boldsymbol{\lambda}_1 - \hat{\beta}\mathbf{T}$ then differentiating the log likelihood with respect to ρ and equating to zero will give

$$\frac{\partial log f}{\partial \rho} = \frac{(t-1)\rho}{(1-\rho^2)} - \frac{t}{2}\frac{\mathbf{e}'\frac{d\Sigma^{-1}}{d\rho}\mathbf{e}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}} = 0, \tag{3.4.5}$$

where

$$\frac{d\Sigma^{-1}}{d\rho} = \frac{1}{(1-\rho^2)^2}\begin{bmatrix} 2\rho & -(1+\rho^2) & 0 & \cdots & 0 \\ -(1+\rho^2) & 4\rho & -(1+\rho^2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 4\rho & -(1+\rho^2) \\ 0 & \cdots & \cdots & -(1+\rho^2) & 2\rho \end{bmatrix},$$

$$\mathbf{e}'\frac{d\Sigma^{-1}}{d\rho}\mathbf{e} = \frac{1}{(1-\rho^2)^2}\left[2\rho e_1^2 + 2\rho e_t^2 - 2(1+\rho^2)\sum_{i=1}^{t-1} e_i e_{i+1} + 4\rho\sum_{i=2}^{t-1} e_i^2\right], \tag{3.4.6}$$

$$\mathbf{e}'\Sigma^{-1}\mathbf{e} = \frac{1}{(1-\rho^2)}\left[e_1^2 + e_t^2 - 2\rho\sum_{i=1}^{t-1} e_i e_{i+1} + (1+\rho^2)\sum_{i=2}^{t-1} e_i^2\right]. \tag{3.4.7}$$

Substituting $\mathbf{e}'\frac{d\Sigma^{-1}}{d\rho}\mathbf{e}$ and $\mathbf{e}'\Sigma^{-1}\mathbf{e}$ in equation 3.4.5 will give

$$(t-1)\rho - t\frac{\left[\rho e_1^2 + \rho e_t^2 - (1+\rho^2)\sum_{i=1}^{t-1} e_i e_{i+1} + 2\rho\sum_{i=2}^{t-1} e_i^2\right]}{\left[e_1^2 + e_t^2 - 2\rho\sum_{i=1}^{t-1} e_i e_{i+1} + (1+\rho^2)\sum_{i=2}^{t-1} e_i^2\right]} = 0. \tag{3.4.8}$$

That is,

$$((t-1)\rho - t\rho)(e_1^2+e_t^2) + (t(1+\rho^2) - 2\rho^2(t-1))\sum_{i=1}^{t-1} e_i e_{i+1} + ((1+\rho^2)(t-1)\rho - 2t\rho)\sum_{i=2}^{t-1} e_i^2 = 0, \tag{3.4.9}$$

$$\rho^3(t-1)\sum_{i=2}^{t-1} e_i^2 + \rho^2(2-t)\sum_{i=1}^{t-1} e_i e_{i+1} - \rho\left[e_1^2 + e_t^2 + (T+1)\sum_{i=2}^{t-1} e_i^2\right] + t\sum_{i=1}^{t-1} e_i e_{i+1} = 0. \tag{3.4.10}$$

The iterative solutions to equation 3.4.3 and 3.4.10 will converge to the maximum likelihood estimates of $\hat{\beta}$ and $\hat{\rho}$. Similarly we can fit correlated linear models on $\mathbf{D}_{1t}$ and the other variables.

To illustrate these procedures we resolved to simulation. First we need to generate the repeated contingency tables for t time periods. For this, we used Mav (2004)'s extension of Sim (1993) algorithm to generate correlated Poisson count data. Also see Chaganty and Mav (2007). The cell frequency for a given contingency table is used as a mean of a Poisson distribution. In our case we used the following contingency table

$$B = \begin{array}{|cccc|}\hline 64 & 94 & 58 & 46 \\ 57 & 94 & 54 & 40 \\ 57 & 105 & 65 & 60 \\ 72 & 141 & 77 & 94 \\ 36 & 97 & 54 & 78 \\ 21 & 71 & 54 & 71 \\ \hline \end{array}$$

to generate repeated contingency tables for time periods $t = 5, 10, 15, 20, 25, 30$. Suppose n_{ijk}, $k = 1, 2, \ldots t$ are the correlated Poisson counts (generated using Sim's algorithm) whose means are changing over time. Mean at time $k = 1$ is given by b_{ij}, the $(i,j)^{th}$ cell frequency of contingency table B. The results based on 500 simulations of fitting a general linear model with correlated errors corresponding to parameter $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{U}_1$, $\mathbf{V}_1$, $\mathbf{U}_2$ and $\mathbf{V}_2$ is shown in Tables 3.6 - 3.13 respectively. The 95^{th}, 90^{th} and 75^{th} percentiles and median of the p-values to test the null hypothesis, $H_o : \beta = 0$, are denoted by P95P, P90P, Q3P, and P50P respectively. Similarly estimates of correlation quantiles were denoted by R95, R90, R3P, and R50. It can be seen in Table 3.6 that the value of P90P is 0 for 25^{th} time period and $R = 0$. Small 90^{th} percentiles of the p-values suggests that when intial Poisson counts are independent then we need at least 25 contingency tables to identify the repeated effect in them. However, as the value of R increases, we need fewer number of repeated contingency tables. Table 3.6 shows that we need at least 10 contingency tables to reject the null hypothesis H_0 when R is non zero. Results of 500 simulations of the first singular value $\boldsymbol{\lambda}_1$ when the mean is not changing over time are shown in Table 3.14. When R is not 0.5 then P15, 15^{th} percentile of the p-values to test the null hypothesis, ranges from 0.051 to 0.143. Hence we can say that 85% of time we accept H_0 at 5% significance level when simulated data does not have time effect. From the simulation results it is quite clear that the first singular value $\boldsymbol{\lambda}_1$ is successful in capturing the repeated effect (mean changing over time) in contingency table.

III.4.1 Performing Correspondence Analysis with repeated contingency table

Let N_i is the i^{th} contingency table, where $i = 1, \ldots, t$. As it is shown in the above section the repeated effect in contingency tables observed over time can be detected by fitting a correlated linear model of the first singular value $\boldsymbol{\lambda}_1$. Hence we can study the given contingency tables, N_i's, by fitting the correlated model of the first singular value $\boldsymbol{\lambda}_1$. If the analysis shows that there is no repeated effect in the contingency tables then we can do correspondence analysis on the combined contingency table $\mathbf{N}$ given by

$$\mathbf{N} = N_1 + N_2 + \ldots + N_t. \tag{3.4.11}$$

If there is any repeated effect in the contingency tables and interest is to see how relationship is changing between two categorical variables over time then it better to perform correspondence analysis on each table separately and interpret the results. But if we want to combine the results and see how the categories at different time periods are related to each other, we can perform correspondence analysis on the contingency table $\mathcal{N}$ which is given by

$$\mathcal{N} = \begin{bmatrix} N_1 & 0 & \ldots & 0 \\ 0 & N_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & N_t \end{bmatrix} \tag{3.4.12}$$

By doing correspondence analysis on contingency table $\mathcal{N}$ we can plot the profiles of different categories and the plot can be used for better understanding of the relationship between the categories of different time periods.

To demonstrate the above method we use the data given in Table 4.12. For our example data let four column categories be denoted by $Y1, Y2, Y3$ and $Y4$ and let six row categories be denoted by $X1, X2, X3, X4, X5$ and $X6$. The output of the correlated linear model on $\boldsymbol{\lambda}_1$ is shown in Table 3.15. Since time has a significant effect on $\boldsymbol{\lambda}_1$ it can be concluded that the contingency tables corresponding to 10 time periods possess the repeated effect. Hence the 10 contingency tables can not be merged together. We therefore use the block diagonal contingency table $\mathcal{N}$ for our analysis.

III.5 Concluding Remarks

In this chapter we demonstrated the use of correspondence analysis for longitudinally observed contingency tables. In order to determine the effect of repeated measure (or the longitudinal effect), we used correlated linear models fitting summaries statistics resulted in performing correspondence analysis (CA) of contingency tables at different time periods on time. Using simulation experiments we determined that the first singular values obtained as a result of correspondence analysis is the best statistical measure of the time effect.

Table 3.1: Socioeconomic Status by Mental Health of Children Data

Parent Socioeconomic Status	Mental Health Status			
	Well	Mild	Moderate	Impaired
1(high)	64	94	58	46
2	57	94	54	40
3	57	105	65	60
4	72	141	77	94
5	36	97	54	78
6(Low)	21	71	54	71

Table 3.2: Mental Health Data: Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.16132	0.02602	43.2013	93.95	93.95
0.03714	0.00138	2.2894	4.98	98.92
0.01726	0.00030	0.4946	1.08	100.00
Total	0.02770	45.9853	100.00	

Table 3.3: Mental Health Data: Canonical Correlations

Canonical correlations ρ	ρ^2	Percent	Cumulative Percent
0.16132	0.02602	93.95	93.95
0.03714	0.00138	4.98	98.92
0.01726	0.00030	1.08	100.00
Total	0.02770	100.00	

Table 3.4: Two Dimensional Coordinates for Socioeconomic Status: (Standardized Form: Mean = 0, SD = 1)

Socioeconomic status	CCA		CA	
	Dim1	Dim2	Dim1	Dim2
1	0.9850	-0.3508	0.9849	0.3499
2	1.0064	-0.1645	1.0065	0.1642
3	0.3432	-0.4228	0.3432	0.4232
4	-0.0143	1.1482	-0.0143	-1.1483
5	-0.8382	1.1855	-0.8382	-1.1850
6	-1.4820	-1.3955	-1.4820	1.3960

Table 3.5: Two Dimensional Coordinates Mental Health Status: (Standardized Form: Mean = 0, SD = 1)

Mental Health Status	CCA		CA	
	Dim1	Dim2	Dim1	Dim2
Well	1.2283	0.3591	1.2282	-0.3587
Mild	0.0992	0.6198	0.0993	-0.6204
Moderate	-0.1158	-1.4914	-0.1158	1.4913
Impaired	-1.2117	0.5125	-1.2117	-0.5122

Table 3.6: First Singular Value λ_1: AR1 Structure Result (Truncated at 2 Decimal Places)

First Singular Value λ_1								
Variable	Correlation $R = 0$				Correlation $R = 0.1$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.85	0.62	0.13	0	0.79	0.12	0.09	0
R95	0.01	0.29	0.29	0.34	0.01	0.42	0.38	0.4
P90P	0.67	0.45	0.06	0	0.61	0.04	0.04	0
R90	-0.08	0.17	0.21	0.26	-0.1	0.26	0.3	0.32
Q3P	0.39	0.15	0.01	0	0.33	0.01	0	0
R3P	-0.33	-0.02	0.07	0.14	-0.37	0.07	0.14	0.2
P50P	0.13	0.04	0	0	0.13	0	0	0
R50	-0.65	-0.24	-0.08	-0.03	-0.63	-0.2	-0.04	0.05
Variable	Correlation $R = 0.2$				Correlation $R = 0.3$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.77	0.03	0	0	0.73	0.03	0	0
R95	0.04	0.42	0.54	0.52	0.02	0.45	0.57	0.72
P90P	0.61	0.01	0	0	0.59	0.01	0	0
R90	-0.06	0.31	0.46	0.44	-0.07	0.34	0.51	0.66
Q3P	0.3	0	0	0	0.28	0	0	0
R3P	-0.34	0.11	0.29	0.34	-0.3	0.19	0.36	0.56
P50P	0.14	0	0	0	0.12	0	0	0
R50	-0.63	-0.12	0.09	0.18	-0.59	-0.08	0.2	0.44
Variable	Correlation $R = 0.4$				Correlation $R = 0.5$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.77	0.06	0	0	0.71	0.04	0	0
R95	0.04	0.51	0.64	0.7	0.04	0.6	0.71	0.79
P90P	0.59	0.02	0	0	0.52	0.01	0	0
R90	-0.08	0.4	0.54	0.64	-0.03	0.51	0.63	0.73
Q3P	0.27	0	0	0	0.23	0	0	0
R3P	-0.29	0.19	0.39	0.54	-0.25	0.33	0.48	0.64
P50P	0.09	0	0	0	0.07	0	0	0
R50	-0.62	-0.04	0.2	0.4	-0.56	0.05	0.32	0.52

Table 3.7: Second Singular Value λ_2: AR(1) Structure Result (Truncated at 2 Decimal Places)

Second Singular Value λ_2								
Variable	Correlation $R = 0$				Correlation $R = 0.1$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.93	0.93	0.94	0.96	0.96	0.93	0.92	0.9
R95	0.03	0.28	0.26	0.28	0.01	0.4	0.36	0.44
P90P	0.87	0.87	0.87	0.9	0.91	0.87	0.87	0.79
R90	-0.08	0.17	0.18	0.22	-0.09	0.29	0.29	0.38
Q3P	0.72	0.71	0.61	0.74	0.71	0.7	0.65	0.52
R3P	-0.29	-0.03	0.05	0.1	-0.35	0.07	0.15	0.26
P50P	0.42	0.39	0.28	0.46	0.43	0.42	0.34	0.28
R50	-0.62	-0.24	-0.09	-0.02	-0.65	-0.19	-0.01	0.12
Variable	Correlation $R = 0.2$				Correlation $R = 0.3$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.94	0.93	0.8	0.92	0.95	0.95	0.95	0.57
R95	-0.02	0.37	0.46	0.45	0.02	0.49	0.5	0.52
P90P	0.88	0.87	0.69	0.84	0.89	0.87	0.87	0.38
R90	-0.1	0.27	0.4	0.39	-0.06	0.37	0.44	0.46
Q3P	0.71	0.67	0.36	0.62	0.69	0.65	0.63	0.12
R3P	-0.3	0.1	0.26	0.28	-0.24	0.2	0.29	0.34
P50P	0.44	0.43	0.12	0.26	0.39	0.37	0.27	0.02
R50	-0.61	-0.12	0.08	0.13	-0.57	-0.02	0.14	0.2
Variable	Correlation $R = 0.4$				Correlation $R = 0.5$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.94	0.94	0.9	0.91	0.94	0.91	0.86	0.85
R95	0.05	0.49	0.51	0.53	0.01	0.5	0.59	0.59
P90P	0.87	0.83	0.82	0.83	0.9	0.82	0.75	0.66
R90	-0.05	0.36	0.43	0.45	-0.08	0.42	0.5	0.52
Q3P	0.7	0.63	0.55	0.55	0.7	0.6	0.43	0.3
R3P	-0.25	0.18	0.27	0.34	-0.25	0.21	0.36	0.42
P50P	0.39	0.35	0.26	0.22	0.4	0.29	0.12	0.05
R50	-0.57	-0.08	0.11	0.21	-0.55	-0.05	0.2	0.28

Table 3.8: First Dimension **D1***: AR(1) Structure Result (Truncated at 2 Decimal Places)*

First Dimension **D1**								
Variable	Correlation $R = 0$				Correlation $R = 0.1$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	1	0.89	1	1	1	1
R95	1	0.99	0.97	0.83	1	1	0.98	0.88
P90P	1	1	1	0.56	1	1	1	1
R90	1	0.99	0.94	0.77	1	1	0.98	0.82
Q3P	0.99	1	0.58	0.05	1	0.96	0.93	0.84
R3P	0.99	0.98	0.81	0.66	0.99	0.99	0.89	0.73
P50P	0.83	0.77	0.01	0	0.91	0.56	0.17	0.15
R50	0.99	0.94	0.6	0.53	0.99	0.99	0.64	0.59
Variable	Correlation $R = 0.2$				Correlation $R = 0.3$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	1	1	1	1	1	1
R95	1	1	1	0.97	1	1	1	1
P90P	1	1	0.98	1	1	1	0.95	0.99
R90	1	1	1	0.95	1	1	1	0.99
Q3P	1	0.99	0.83	0.93	0.99	0.96	0.69	0.7
R3P	0.99	0.99	0.99	0.88	1	1	1	0.89
P50P	0.91	0.82	0.26	0.16	0.85	0.62	0.24	0.08
R50	0.99	0.99	0.79	0.71	0.99	0.99	0.77	0.72
Variable	Correlation $R = 0.4$				Correlation $R = 0.5$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	1	1	1	1	1	1
R95	1	1	1	0.93	1	1	1	0.99
P90P	1	1	0.99	1	1	1	1	1
R90	1	1	0.99	0.88	1	1	1	0.93
Q3P	0.99	0.98	0.85	0.82	0.99	0.96	0.78	0.89
R3P	1	0.99	0.82	0.75	1	1	0.9	0.85
P50P	0.79	0.58	0.17	0.17	0.78	0.59	0.25	0.2
R50	0.99	0.85	0.59	0.54	0.99	0.99	0.68	0.65

Table 3.9: Second Dimension **D2***: AR(1) Structure Result (Truncated at 2 Decimal Places)*

Second Dimension **D2**								
Variable	Correlation $R = 0$				Correlation $R = 0.1$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	0.98	0.2	0.01	1	1	0.26	0
R95	0.97	0.86	0.51	0.38	0.98	0.97	0.8	0.74
P90P	0.98	0.93	0.07	0	1	1	0.06	0
R90	0.96	0.82	0.42	0.32	0.98	0.97	0.74	0.69
Q3P	0.74	0.4	0.01	0	0.96	0.97	0	0
R3P	0.92	0.61	0.32	0.26	0.95	0.93	0.61	0.61
P50P	0.16	0	0	0	0.49	0.47	0	0
R50	0.81	0.36	0.21	0.17	0.9	0.86	0.48	0.49
Variable	Correlation $R = 0.2$				Correlation $R = 0.3$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	0.93	0.03	1	1	0.89	0.33
R95	0.96	0.93	0.91	0.77	0.97	0.95	0.91	0.88
P90P	0.98	0.96	0.55	0	1	0.99	0.28	0.07
R90	0.95	0.88	0.87	0.72	0.96	0.93	0.85	0.83
Q3P	0.78	0.34	0.02	0	0.92	0.68	0.01	0
R3P	0.86	0.76	0.76	0.62	0.92	0.84	0.73	0.75
P50P	0.2	0.01	0	0	0.42	0.08	0	0
R50	0.57	0.54	0.6	0.52	0.74	0.56	0.56	0.62
Variable	Correlation $R = 0.4$				Correlation $R = 0.5$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	1	0.42	1	1	1	1
R95	0.98	0.97	0.93	0.87	0.99	0.98	0.97	0.96
P90P	1	1	0.88	0.12	1	1	1	0.97
R90	0.97	0.95	0.9	0.82	0.98	0.97	0.96	0.95
Q3P	0.9	0.87	0.13	0	0.94	0.96	0.87	0.44
R3P	0.94	0.9	0.8	0.74	0.96	0.94	0.92	0.89
P50P	0.35	0.13	0	0	0.35	0.45	0.08	0.01
R50	0.81	0.71	0.68	0.64	0.89	0.87	0.82	0.81

Table 3.10: First Principal Axis for Row **U1***: AR(1) Structure Result (Truncated at 2 Decimal Places)*

First Principal Axis for Row **U1**								
Variable	Correlation $R = 0$				Correlation $R = 0.1$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	0.99	0.98	1	1	1	0.98
R95	0.99	0.6	0.43	0.36	1	0.74	0.55	0.43
P90P	1	1	0.98	0.95	1	1	0.99	0.96
R90	0.86	0.52	0.33	0.28	0.99	0.66	0.46	0.36
Q3P	0.88	0.96	0.91	0.82	0.89	0.98	0.97	0.89
R3P	0.81	0.28	0.2	0.17	0.85	0.47	0.32	0.25
P50P	0.67	0.66	0.56	0.41	0.66	0.72	0.78	0.65
R50	0.26	0.08	0.05	0.06	0.26	0.2	0.13	0.13
Variable	Correlation $R = 0.2$				Correlation $R = 0.3$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	1	1	1	1	1	1
R95	1	0.77	0.57	0.49	1	0.88	0.65	0.57
P90P	1	1	1	0.99	1	1	1	1
R90	0.99	0.7	0.54	0.43	1	0.71	0.56	0.5
Q3P	0.94	0.99	0.99	0.96	0.98	0.99	0.99	0.99
R3P	0.81	0.49	0.34	0.31	0.86	0.52	0.41	0.37
P50P	0.67	0.81	0.78	0.77	0.67	0.72	0.77	0.84
R50	0.31	0.21	0.16	0.19	0.35	0.26	0.23	0.23
Variable	Correlation $R = 0.4$				Correlation $R = 0.5$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	1	1	1	1	1	0.99
R95	1	0.91	0.66	0.58	1	0.99	0.77	0.68
P90P	1	1	1	0.99	1	1	1	0.99
R90	0.99	0.72	0.56	0.51	1	0.92	0.67	0.58
Q3P	0.95	0.96	0.97	0.93	0.95	0.96	0.96	0.94
R3P	0.86	0.52	0.44	0.39	0.98	0.66	0.52	0.49
P50P	0.66	0.61	0.79	0.67	0.66	0.65	0.67	0.66
R50	0.36	0.27	0.27	0.26	0.36	0.38	0.34	0.36

Table 3.11: First Principal Axis for Column **V1***: AR(1) Structure Result (Truncated at 2 Decimal Places)*

First Principal Axis for Column **V1**								
Variable	Correlation $R = 0$				Correlation $R = 0.1$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	1	0.99	1	1	0.99	0.96
R95	1	0.62	0.45	0.38	1	0.75	0.57	0.43
P90P	1	1	0.98	0.97	1	1	0.99	0.93
R90	0.87	0.54	0.34	0.3	1	0.68	0.46	0.36
Q3P	0.85	0.94	0.92	0.85	0.85	0.96	0.95	0.84
R3P	0.83	0.29	0.2	0.18	0.87	0.49	0.32	0.25
P50P	0.62	0.66	0.55	0.42	0.62	0.71	0.77	0.63
R50	0.28	0.09	0.06	0.05	0.28	0.22	0.13	0.13
Variable	Correlation $R = 0.2$				Correlation $R = 0.3$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	1	0.99	1	1	1	1
R95	1	0.79	0.58	0.5	1	0.89	0.66	0.58
P90P	1	1	1	0.98	1	1	1	0.99
R90	1	0.72	0.55	0.44	1	0.73	0.57	0.51
Q3P	0.87	0.97	0.98	0.93	0.95	0.96	0.96	0.97
R3P	0.83	0.51	0.35	0.31	0.87	0.54	0.42	0.37
P50P	0.62	0.76	0.76	0.71	0.62	0.69	0.72	0.78
R50	0.36	0.23	0.17	0.18	0.38	0.28	0.23	0.23
Variable	Correlation $R = 0.4$				Correlation $R = 0.5$			
TimeP	5	10	20	30	5	10	20	30
P95P	1	1	1	0.99	1	1	1	0.99
R95	1	0.92	0.68	0.58	1	1	0.78	0.69
P90P	1	0.99	0.99	0.98	1	0.99	0.99	0.99
R90	1	0.73	0.57	0.51	1	0.92	0.68	0.59
Q3P	0.9	0.94	0.95	0.91	0.9	0.94	0.94	0.94
R3P	0.87	0.54	0.45	0.4	0.99	0.68	0.54	0.5
P50P	0.61	0.58	0.77	0.64	0.61	0.6	0.65	0.66
R50	0.38	0.28	0.27	0.27	0.38	0.4	0.35	0.36

Table 3.12: Second Principal Axis for Row **U2***: AR(1) Structure Result (Truncated at 2 Decimal Places)*

Second Principal Axis for Row **U2**								
Variable	Correlation $R = 0$				Correlation $R = 0.1$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.89	0.89	0.18	0.01	0.99	0.99	0.54	0.02
R95	0.99	0.91	0.57	0.44	0.99	0.99	0.89	0.78
P90P	0.8	0.81	0.09	0.01	0.99	0.98	0.31	0
R90	0.99	0.87	0.5	0.39	0.99	0.99	0.82	0.7
Q3P	0.63	0.57	0.03	0	0.96	0.93	0.09	0
R3P	0.98	0.74	0.41	0.32	0.99	0.98	0.7	0.61
P50P	0.47	0.17	0.01	0	0.82	0.68	0.02	0
R50	0.93	0.46	0.31	0.24	0.98	0.95	0.56	0.49
Variable	Correlation $R = 0.2$				Correlation $R = 0.3$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.96	0.89	0.77	0.07	0.98	0.95	0.69	0.27
R95	0.97	0.94	0.95	0.82	0.99	0.97	0.94	0.91
P90P	0.93	0.77	0.64	0.02	0.96	0.9	0.55	0.14
R90	0.96	0.92	0.92	0.75	0.98	0.96	0.91	0.87
Q3P	0.79	0.49	0.25	0	0.88	0.78	0.23	0.03
R3P	0.93	0.84	0.83	0.65	0.97	0.93	0.84	0.82
P50P	0.56	0.19	0.02	0	0.66	0.56	0.05	0
R50	0.82	0.68	0.68	0.56	0.91	0.86	0.73	0.72
Variable	Correlation $R = 0.4$				Correlation $R = 0.5$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.99	0.98	0.87	0.35	0.99	0.99	0.96	0.78
R95	0.99	0.99	0.97	0.9	0.99	0.99	0.99	0.97
P90P	0.98	0.97	0.75	0.12	0.98	0.98	0.9	0.62
R90	0.99	0.99	0.94	0.83	0.99	0.99	0.98	0.95
Q3P	0.94	0.89	0.27	0.01	0.93	0.94	0.55	0.15
R3P	0.98	0.97	0.83	0.73	0.99	0.98	0.93	0.87
P50P	0.7	0.41	0.02	0	0.68	0.51	0.1	0.01
R50	0.93	0.76	0.68	0.64	0.98	0.92	0.77	0.76

Table 3.13: Second Principal Axis for Column **V2***: AR(1) Structure Result (Truncated at 2 Decimal Places)*

Second Principal Axis for Column **V2**								
Variable	Correlation $R = 0$				Correlation $R = 0.1$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.99	0.93	0.48	0.11	0.99	0.98	0.77	0.26
R95	1	0.99	0.74	0.58	1	1	0.97	0.89
P90P	0.97	0.89	0.26	0.04	0.97	0.94	0.63	0.09
R90	1	0.98	0.66	0.51	1	1	0.93	0.83
Q3P	0.8	0.67	0.09	0.01	0.9	0.8	0.22	0.01
R3P	1	0.92	0.52	0.42	1	0.99	0.77	0.71
P50P	0.45	0.3	0.01	0	0.59	0.43	0.05	0
R50	0.99	0.58	0.39	0.33	0.99	0.96	0.64	0.58
Variable	Correlation $R = 0.2$				Correlation $R = 0.3$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.96	0.88	0.86	0.49	0.96	0.9	0.67	0.53
R95	0.99	0.97	0.97	0.93	0.99	0.97	0.95	0.95
P90P	0.93	0.84	0.8	0.23	0.94	0.83	0.6	0.38
R90	0.98	0.95	0.96	0.88	0.98	0.96	0.94	0.92
Q3P	0.77	0.68	0.59	0.04	0.84	0.71	0.37	0.17
R3P	0.95	0.92	0.92	0.75	0.96	0.93	0.88	0.86
P50P	0.6	0.48	0.18	0	0.66	0.5	0.09	0.02
R50	0.84	0.81	0.8	0.63	0.89	0.83	0.76	0.76
Variable	Correlation $R = 0.4$				Correlation $R = 0.5$			
TimeP	5	10	20	30	5	10	20	30
P95P	0.98	0.95	0.8	0.57	0.99	0.96	0.86	0.81
R95	1	1	0.98	0.95	1	1	0.99	0.99
P90P	0.97	0.92	0.69	0.29	0.98	0.91	0.78	0.66
R90	1	0.99	0.96	0.89	1	1	0.99	0.98
Q3P	0.91	0.75	0.47	0.07	0.86	0.71	0.56	0.43
R3P	0.99	0.98	0.89	0.8	1	0.99	0.96	0.93
P50P	0.66	0.4	0.08	0.01	0.51	0.37	0.24	0.07
R50	0.94	0.82	0.73	0.68	0.99	0.94	0.83	0.82

Table 3.14: First Singular Value λ_1, Mean is not Changing Over Time: AR(1) Structure Result

First Singular Value λ_1								
Variable	Correlation $R = 0$				Correlation $R = 0.1$			
TimeP	10	15	20	25	10	15	20	25
P10	0.082	0.115	0.101	0.086	0.061	0.084	0.092	0.071
P15	0.140	0.149	0.148	0.139	0.110	0.129	0.135	0.121
P20	0.181	0.195	0.189	0.190	0.144	0.171	0.174	0.169
R10	-0.622	-0.448	-0.380	-0.325	-0.587	-0.400	-0.321	-0.270
R15	-0.550	-0.398	-0.324	-0.291	-0.508	-0.352	-0.257	-0.210
R20	-0.502	-0.347	-0.281	-0.254	-0.451	-0.286	-0.209	-0.174
Variable	Correlation $R = 0.2$				Correlation $R = 0.3$			
TimeP	10	15	20	25	10	15	20	25
P10	0.050	0.061	0.055	0.044	0.037	0.042	0.031	0.031
P15	0.083	0.092	0.090	0.093	0.074	0.060	0.056	0.064
P20	0.122	0.114	0.128	0.140	0.103	0.102	0.090	0.093
R10	-0.523	-0.340	-0.237	-0.168	-0.445	-0.270	-0.192	-0.090
R15	-0.438	-0.264	-0.178	-0.124	-0.386	-0.197	-0.086	-0.047
R20	-0.372	-0.208	-0.135	-0.091	-0.314	-0.133	-0.041	-0.009
Variable	Correlation $R = 0.4$				Correlation $R = 0.5$			
TimeP	10	15	20	25	10	15	20	25
P10	0.036	0.039	0.029	0.026	0.031	0.024	0.019	0.010
P15	0.071	0.059	0.051	0.051	0.048	0.054	0.030	0.028
P20	0.097	0.093	0.079	0.075	0.066	0.083	0.051	0.047
R10	-0.477	-0.240	-0.132	-0.068	-0.371	-0.127	-0.022	0.081
R15	-0.387	-0.181	-0.087	-0.025	-0.282	-0.079	0.052	0.137
R20	-0.324	-0.123	-0.046	0.018	-0.230	-0.013	0.101	0.179

Table 3.15: Simulated Contingency Table Example : Repeated Effect

Effect	Estimate	Standard Error	DF	t Value	$Pr > \|t\|$
Intercept	0.1660	0.006252	8	26.55	$< .0001$
t	-0.00454	0.001008	8	-4.50	0.0020

CHAPTER IV

CANONICAL CORRESPONDENCE ANALYSIS

IV.1 Introduction

Canonical correspondence analysis (CCPA) is a multivariate data analysis technique which was introduced by Ter Braak (1986) to relate community composition to known variation in the environment. Problems in community ecology often require the determination of species-environment relationship from community composition data and associated habitat measurements. Typical data for such problems consist of the two sets, one the abundance of a number of species at a series of sites, and another data set on a number of environmental variables measured at the same sites. A site here is the basic sampling unit, separated in space or time from other sites. By treating species-abundance data over different sites as a contingency table, correspondence analysis (CA) can be performed to graphically represent these data. Such graphical displays can be helpful in identifying the sites that have a maximum abundance of a certain species. Frequently ecologists are also interested in determining the relationship between the environmental variables favorable for the growth of certain species. Basically the idea is to represent the relationship between the environmental variables and species graphically. This analysis is named as *Canonical correspondence analysis* (CCPA). In the literature this analysis is also abbreviated as CCA, but to avoid confusion with canonical correlation analysis (CCA) we will denote this here by CCPA. Ter Braak (1986) has developed a Fortran program, named CANOCO, to perform this analysis. Hegde and Naik (1999) developed a SAS program to perform the same analysis. Also see Khattree and Naik (2000) for a review and analysis of CCPA.

It is clear, like in the previous two chapters, that here also the basic problem is to study the relationship between two sets of variables. While in Chapter 2 we study the relationship between two sets of quantitative variables, in Chapter 3 we studied the relationship between two sets of qualitative variables. However, here the interest is in studying the relationship between a set of qualitative variables and another set of quantitative variables. In the next section, in order to introduce the notation and background we provide Ter Braak's formulation of CCPA. In Section

3 we will show how CCPA can be performed using canonical correlation analysis of certain variance covariance matrices. In Section 4 we will provide a theoretical basis to matrix formulation of CCPA. In Section 4 we will also provide canonical correspondence analysis to longitudinally observed data.

IV.2 Canonical Correspondence Analysis

In this section we present the details used by Ter Braak (1986) to explain canonical correspondence analysis (CCPA). Suppose a survey of n sites lists the occurrences (presence as 1, absence as 0) of m species and the values of q environmental variables ($q < n$). Let y_{il} represents the abundance of the l^{th} species at the i^{th} site, where $i = 1, \ldots, n$ and $l = 1, \ldots, m$ and z_{ij}, $j = 1, \ldots, q$ is the value of the j^{th} environmental variable at the i^{th} site. Let $\mathbf{Y}_{n\times m} = (y_{il})$ denote the n by m matrix of species abundance.

It is assumed that y_{il} has independent Poisson distribution with mean m_{il}. Canonical correspondence analysis (CCPA) can be considered as a two step method. The first step is to summarize the main variation in the species abundance data by ordination. The method of Gaussian ordination as described by Gauch, Chase and Whittaker (1974) does this by constructing an axis such that the species optimally fit the Gaussian response curve,

$$m_{il} = c_l e^{\frac{-(x_i - \mu_l)^2}{2\sigma_l^2}}, \tag{4.2.1}$$

along the constructed axis. Here m_{il} is the expected value of y_{il} at site i that has site score x_i, which is usually an unknown linear combination of the environmental variables, on the ordination axis; μ_l can be interpreted as the value of x_i resulting in maximum abundance for the l^{th} species; c_l can be interpreted as the value of maximum mean abundance and σ_l as an index of the tolerance, a measure of ecological amplitude. It is shown by the Ter Braak (1985) that correspondence analysis (CA) approximates the maximum likelihood solution of Gaussian ordination, if the sampling distribution of species abundances is Poisson and if these assumptions made, that *The species' tolerances σ_l are all equal to σ; The species' maxima c_l are all equal to c; The species' optima μ_l are homogeneously distributed over an interval I_1*

that is large compared to σ; and The site scores x_i are homogeneously distributed over an interval I_2 that contains I_1.

The second step of canonical correspondence analysis (CCPA) is to relate the ordination axis to the environmental variables by doing multiple regression of the site scores on the environmental variables. Before doing multiple regression, environmental variables z_{ij} are standardized such that their weighted means over all sites are zero, and the corresponding weighted standard deviations are all one. Hence suppose

$$\sum_i w_i z_{ij} = 0 \text{ and } \sum_i w_i z_{ij}^2 = 1,$$

where $w_i = \frac{y_{i.}}{y_{..}}$, $y_{i.} = \sum_l y_{il}$ and $y_{..} = \sum_i \sum_l y_{il}$ and denote the n by q matrix of these standardize environmental variables by $\mathbf{Z} = (z_{ij})$. Then the multiple regression of site scores on environmental variables is given by

$$x_i = \beta_1 z_{i1} + \ldots + \beta_q z_{iq} = \mathbf{z}_i' \boldsymbol{\beta}. \tag{4.2.2}$$

The main objective of canonical correspondence analysis (CCPA) is to estimate the vectors of unknown parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)'$ and $\boldsymbol{\beta}$. This can be done by simultaneously estimating the species optima and regression coefficient by equation 4.2.1 and 4.2.2. Estimating equations in matrix form for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ are given by

$$\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12} - \lambda\mathbf{I} = \boldsymbol{\beta} \tag{4.2.3}$$

$$\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} - \lambda\mathbf{I} = \boldsymbol{\mu}, \tag{4.2.4}$$

where $\mathbf{S}_{21} = \mathbf{Z}'\mathbf{Y}$, $\mathbf{S}_{12} = \mathbf{Y}'\mathbf{Z}$, $\mathbf{S}_{11} = diag(y_{.1}, \ldots, y_{.m})$ and $\mathbf{S}_{22} = \mathbf{Z}'\mathbf{D}\mathbf{Z}$, with $\mathbf{D} = diag(y_{1.}, \ldots, y_{n.})$. The solution to equations 4.2.3 and 4.2.4 are obtained by singular value decomposition of the matrix

$$\mathbf{W} = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}.$$

Singular value decomposition of matrix $\mathbf{W}$ is given by

$$\mathbf{W} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'. \tag{4.2.5}$$

Hence r (rank of $\mathbf{W}$) solutions of equations 4.2.3 and 4.2.4 are given by the matrices:

$$\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1 : \ldots : \hat{\boldsymbol{\beta}}_r) = \mathbf{S}_{22}^{-1/2}\mathbf{V}$$

and

$$\hat{\mathbf{M}} = (\hat{\boldsymbol{\mu}}_1 : \ldots : \hat{\boldsymbol{\mu}}_r) = \mathbf{S}_{11}^{-1/2}\mathbf{U}.$$

The site score matrix and species scores matrix are given by

$$\mathbf{X} = \mathbf{Z}\hat{\mathbf{B}} \tag{4.2.6}$$

and

$$\hat{\mathbf{M}} = \mathbf{S}_{11}^{-1/2}\mathbf{Y}'\mathbf{X}\boldsymbol{\Lambda}^{-1} \tag{4.2.7}$$

respectively.

More detailed mathematical explanations and geometrical interpretations of canonical correspondence analysis (CCPA) can be found in Hegde and Naik (2006, Preprint). To illustrate the method we use an example in the following subsection.

IV.2.1 Hunting Spider Example

Although data description and analysis are given in Khattree and Naik (2000), for completeness sake we will describe the same here. The data considered in this example is from Ter Braak (1986), Table 3, and were originally adapted from Van der Aart and Smeek-Enseink (1975) after transformation. Data consist of abundance of 12 species of hunting spider at 28 sites, representing pitfall traps, caught in pitfall traps over a period of 60 weeks, along with measurements on six environmental variables, namely percentage of soil dry mass, percentage cover of bare sand, percentage cover of fallen leaves and twigs, percentage cover of the herb layer and reflection of the soil surface with cloudless sky. The square root transformation was performed on the species abundance and a logarithmic transformation was performed on the environmental variables. Only the integer part of the square root transformed abundance were considered. A value of 9 for species abundance indicates the number of individuals of the species found is greater than or equal to eighty one. Further, the range of each transformed environmental variable was divided into 10 equal categories denoted by 0-9 and these numbers were used as the data corresponding to the environmental variables. An objective of the study was to determine the distribution of these 12 species of hunting spiders in a Dutch dune area in relation to the environmental variables. Species abundance data and environmental data at 28 pitfall

traps are shown in Table 4.2 and 4.3 respectively. The fundamental matrix **W** for the hunting spider data is given by

$$\mathbf{W} = \begin{bmatrix} 0.0905928 & 0.0221424 & -0.018828 & -0.0339 & -0.042086 & 0.043711 \\ -0.07908 & -0.120468 & -0.122306 & -0.144538 & 0.2624115 & -0.147858 \\ 0.0985968 & 0.0062076 & -0.048164 & -0.124552 & 0.0042143 & 0.02378 \\ 0.118231 & -0.006221 & -0.090492 & -0.013082 & -0.065483 & 0.0897482 \\ 0.1194795 & -0.077528 & -0.022274 & -0.016324 & -0.095579 & 0.1384742 \\ 0.0441676 & -0.030179 & -0.047171 & -0.019193 & -0.009464 & 0.1225125 \\ 0.0698714 & -0.078222 & -0.084326 & -0.095609 & 0.1093381 & -0.050272 \\ 0.05573 & -0.056263 & -0.0273 & 0.0460913 & 0.0662339 & 0.0193313 \\ -0.01269 & -0.012869 & 0.2247585 & 0.1192219 & -0.072703 & 0.002098 \\ -0.229019 & 0.1024604 & 0.155209 & 0.1450659 & -0.064858 & -0.030001 \\ -0.257685 & 0.2716802 & 0.0779809 & 0.1071344 & -0.088528 & -0.151806 \\ -0.241603 & 0.2128121 & 0.0113994 & 0.1153476 & -0.064448 & -0.166911 \end{bmatrix}$$

The canonical correlation resulted by canonical correspondence analysis (CCPA) and their contribution towards the variance explained is shown in Table 4.1.

Corresponding site and species scores resulted by canonical correspondence analysis (CCPA) are shown in Tables 4.4 and 4.5. The Biplot graphical display for the hunting spider data is given by Figure 4.1. The graphical display suggests that the species Alop-fabr (Al-f) and Arct-peri (Ar-p) were mainly found in habitats with higher percentage of sand (BARE-SAND). The species Arct-lute (Ar-l), Pard-pull (Pa-p), Pard-mont (Pa-m), Pard-nigr (Pa-n) and Aulo-albi (Au-a) are found in habitats with well developed herb layers (COVER-HERBS). Only the species Pard-lugu (Pa-l) is found in the habitats with fallen twigs and leaves represented by the variable FALLEN-TWIGS in the graph.

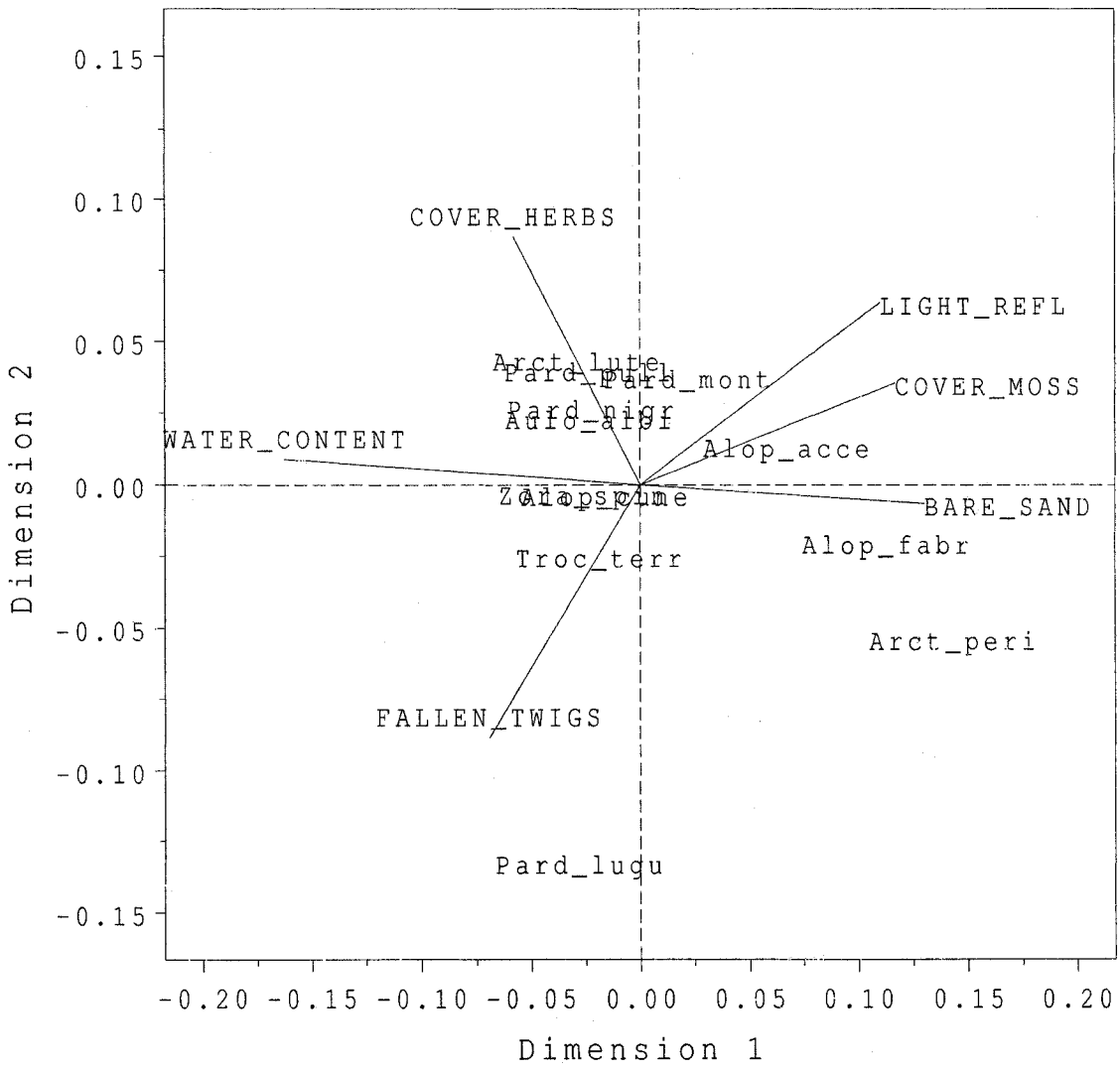

Figure 4.1: Biplot of Hunting Spider Data.

Table 4.1: Hunter Spider: Canonical Correlations

Cannonical Correlation	Eigenvalue	Percent	Cumulative %
0.7341518	0.538978865	61.88304447	61.88304447
0.473724	0.224414428	25.76622003	87.64926449
0.2698241	0.072805045	8.359136364	96.00840086
0.1407824	0.019819684	2.275603876	98.28400473
0.1063434	0.011308919	1.298437407	99.58244214
0.0603057	0.003636777	0.417557858	100

Table 4.2: Hunting Spider Species Abundance Data

Sites	Ar-I	Pa-L	Zo-s	Pa-n	Pa-p	Au-a	Tr-t	Al-c	Pa-m	Al-a	Al-f	Ar-p
1	0	2	1	0	0	0	5	0	0	0	0	0
2	0	3	1	1	0	0	4	1	0	0	0	0
3	0	3	1	0	0	0	4	1	0	0	0	0
4	0	2	2	1	0	0	5	1	0	0	0	0
5	0	1	1	0	0	0	4	0	0	0	0	0
6	0	2	0	0	0	0	5	1	0	0	0	0
7	0	1	3	3	6	5	8	1	1	0	0	0
8	0	7	1	1	1	2	5	3	1	0	0	0
9	0	4	1	0	1	0	4	1	1	0	0	0
10	1	1	4	9	8	3	9	4	1	1	0	0
11	2	0	5	5	4	2	7	2	3	0	0	0
12	1	1	5	3	8	2	9	1	3	0	0	0
13	1	1	5	5	9	4	9	2	2	1	0	0
14	3	1	4	9	9	4	9	2	5	1	0	0
15	1	1	4	7	8	4	9	6	4	1	1	0
16	1	1	1	4	6	3	8	4	5	3	1	0
17	0	0	2	3	6	2	7	3	7	5	0	0
18	0	0	0	1	1	0	1	1	5	1	0	0
19	0	0	0	1	2	0	3	3	9	4	0	0
20	0	1	2	2	0	1	4	1	3	3	3	0
21	0	0	0	0	1	1	2	1	9	3	1	0
22	0	0	0	0	0	0	1	0	4	1	1	0
23	0	0	0	0	0	0	1	0	2	3	3	1
24	0	1	0	0	0	0	1	0	2	4	3	2
25	0	0	0	0	0	0	1	0	1	2	4	1
26	0	0	0	0	0	0	0	0	1	5	3	2
27	0	0	0	0	0	0	0	0	1	3	4	2
28	0	0	0	0	0	0	1	0	0	1	2	4

Table 4.3: Hunting Spider Data Observed on 6 Environmental Variables for 28 Sites

Sites	Water Content	Bare Sand	Cover Moss	Light Reft	Fallen Twigs	Cover Herbs
1	9	0	1	1	9	5
2	7	0	3	0	9	2
3	8	0	1	0	9	0
4	8	0	1	0	9	0
5	9	0	1	2	9	5
6	8	0	0	2	9	5
7	8	0	2	3	3	9
8	6	0	2	1	9	6
9	7	0	1	0	9	2
10	8	0	0	5	0	9
11	9	5	5	1	7	6
12	8	0	4	2	0	9
13	6	0	5	6	0	9
14	8	0	1	5	0	9
15	9	3	1	7	3	9
16	6	0	5	8	0	9
17	5	0	7	8	0	9
18	5	0	9	7	0	6
19	6	0	8	8	0	8
20	3	7	2	5	0	8
21	4	0	9	8	0	7
22	4	8	7	8	0	5
23	0	7	8	8	0	6
24	0	6	9	9	0	6
25	1	7	9	8	0	0
26	0	5	8	8	0	6
27	2	7	9	9	0	5
28	0	9	4	9	0	2

Table 4.4: CCPA: Site Scores (Standardized Form: Mean = 0, SD = 1)

Site	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6
1	-1.1158	-0.9066	0.3801	-0.0755	0.2837	0.8169
2	-0.6289	-1.4222	0.6536	0.8009	-0.2028	-0.4257
3	-0.6153	-1.6232	-0.0219	1.2911	1.2458	-1.4878
4	-0.6153	-1.6232	-0.0219	1.2911	1.2458	-1.4878
5	-1.0496	-0.8916	0.5568	-0.3084	0.8261	0.9413
6	-0.9283	-1.3203	0.3483	-1.0637	0.5093	0.7142
7	-0.9765	0.3872	-0.2516	-0.4947	-0.9478	0.2506
8	-0.7874	-1.4951	0.8415	-1.4583	-1.5292	1.1536
9	-0.6543	-1.7284	0.1375	0.2005	0.0344	-0.7245
10	-0.7663	0.6338	-1.1703	-0.8247	0.2691	-1.0837
11	-0.6306	0.5598	-0.4668	2.0894	-0.6019	2.4838
12	-0.9141	1.2013	-0.6685	1.0749	-1.8326	-0.8595
13	-0.3684	0.8630	0.3955	-0.4669	-0.6524	-0.3679
14	-0.7536	0.7869	-0.9123	-0.5245	0.1505	-0.9344
15	-0.5872	0.8218	-0.7100	-0.7970	1.8040	1.5797
16	-0.2360	0.8929	0.7489	-0.9329	0.4325	-0.1191
17	-0.0765	0.9235	1.3145	-0.7876	-0.2402	0.1020
18	0.1423	0.9592	1.4890	0.9990	0.1442	-0.9852
19	-0.1114	1.2671	1.4681	0.2854	0.4646	-0.0914
20	0.6491	0.0651	-2.4635	-1.0747	-1.6915	0.5129
21	0.2561	0.7837	1.7701	-0.0068	-0.1368	-0.5180
22	1.1272	0.9796	-1.1434	1.2939	0.9515	0.7503
23	1.4993	0.0314	-0.3465	-0.7022	-1.3062	0.7082
24	1.4878	0.1154	0.3742	-0.7932	-0.8917	0.6810
25	1.8971	-0.0511	-0.4676	1.9595	1.3386	-1.5877
26	1.3184	-0.1367	0.2254	-1.0187	-1.3249	0.1062
27	1.3965	0.6655	-0.0657	0.5929	0.3765	0.7170
28	2.0416	-0.7388	-1.9935	-0.5486	1.2814	-0.8449

Table 4.5: CCPA: Species Scores (Standardized Form: Mean = 0, SD = 1)

Species	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6
Ar-I	-0.6913	0.9620	-1.0973	1.5969	0.1113	1.1018
Pa-L	-0.6762	-2.5548	1.0037	-0.3359	-0.7031	-0.5241
Zo-s	-0.6513	0.0298	-0.5823	1.4449	-1.0598	0.5209
Pa-n	-0.5832	0.6254	-0.6738	-0.4404	0.6122	-0.4176
Pa-p	-0.6035	0.8832	0.0818	-0.4277	-0.4371	-1.3998
Au-a	-0.5995	0.5546	-0.0967	-1.5129	-1.1171	1.1684
Tr-t	-0.5220	-0.4104	0.2618	0.3843	0.2883	-0.2868
Al-c	-0.4805	0.0185	0.7858	-0.8298	1.8411	1.2281
Pa-m	0.1415	0.8376	1.6617	1.0022	0.2425	-0.2734
Al-a	0.8860	0.3549	1.1386	-0.9710	-1.1417	0.1074
Al-f	1.6397	-0.3144	-0.9652	0.6755	-0.2649	0.6748
Ar-p	2.1401	-0.9863	-1.5180	-0.5861	1.6282	-1.8998

IV.3 Canonical Correspondence Analysis (CCPA) as Canonical Correlation Analysis (CCA)

In this section we will discuss the connection between canonical correspondence analysis and canonical correlation analysis. This connection between CCPA and CCA is helpful in providing theoretical insight to CCPA. We will show how to generate all the results generated by canonical correspondence analysis by using canonical correlation analysis. As we mentioned earlier, CCA is an analysis of two sets of quantitative variables, where as, CCPA is an analysis of two sets of data of which one set is of qualitative in nature (e.g. species abundance data: m different species observed at n different sites) and the other one is quantitative (e.g. data on environmental variables: q different environmental variables observed at n sites). To perform CCA on these data, we first create a matrix of indicator variables indicating in which category of species each of the N individuals from the species abundance matrix $\mathbf{Y}_{n\times m}$ belong. Then we create a large matrix $\mathbf{H}_{N\times(m+q)}$ by augmenting the environmental data matrix $\mathbf{Z}_{n\times q}$ with the indicator matrix. Here $N = \sum_{ij} n_{ij} = n_{..}$. Then we can calculate the canonical scores for different sites, called site scores, based on canonical coefficients of environmental variables. Next based on the site scores we will calculate the species scores. Finally, using the other information provided by CCA we will be able to create a graphical display of species abundance data and environmental data. We use the same Hunting Spider example for illustration.

IV.3.1 Hunting Spider Example

Let y_{il} represents the abundance of the l^{th} species at the i^{th} site, where $i = 1, \ldots, n$ and $l = 1, \ldots, m$ and z_{ij}, $j = 1, \ldots, q$ is the value of the j^{th} environmental variable at the i^{th} site. Let $\mathbf{N}_{n\times m} = (y_{il})$ denote the n by m matrix of species abundance and $\mathbf{Z}_{n\times q} = (z_{ij})$ denote the n by q data matrix of environmental data. Using the species abundance matrix $\mathbf{N}_{n\times m}$ we create an indicator variable matrix, indicating in which category of species each of N individuals belong, and augment that matrix with the environmental data matrix $\mathbf{Z}_{n\times q}$. This would create an $N \times (m+q)$ matrix, $\mathbf{H}_{N\times(m+q)}$, where $N = \sum_{ij} n_{ij} = n_{..}$. A canonical correlation analysis is performed on this matrix by taking the indicator variable matrix as the data on one set of variables and the data on the environmental variables repeated so many times as the second

data set. For example, if $\mathbf{N}_{2\times 2} = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{Z}_{2\times 3} = \begin{bmatrix} 9 & 2 & 5.6 \\ 6.5 & 1 & 6.1 \end{bmatrix}$ then we get

$$\mathbf{H}_{6\times 5} = \begin{bmatrix} 1 & 0 & 9 & 2 & 5.6 \\ 1 & 0 & 9 & 2 & 5.6 \\ 0 & 1 & 9 & 2 & 5.6 \\ 0 & 1 & 9 & 2 & 5.6 \\ 0 & 1 & 9 & 2 & 5.6 \\ 0 & 1 & 6.5 & 1 & 6.1 \end{bmatrix}.$$

To perform CCA we use the estimated variance covariance matrices of $\mathbf{y}$, $\mathbf{z}$, and the covariance between $\mathbf{y}$ and $\mathbf{z}$. The estimated variance covariance matrices given in blocks

$$D\begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{bmatrix}$$

are obtained from absence/presence matrix, $\mathbf{H}_{N\times(m+q)}$. They are given by

$$\hat{\Sigma}_{yy} = \begin{bmatrix} \hat{p}_{.1}(1-\hat{p}_{.1}) & -\hat{p}_{.1}\hat{p}_{.2} & \cdots & -\hat{p}_{.1}\hat{p}_{.m} \\ -\hat{p}_{.1}\hat{p}_{.2} & \hat{p}_{.2}(1-\hat{p}_{.2}) & \cdots & -\hat{p}_{.2}\hat{p}_{.m} \\ \vdots & \vdots & \vdots & \vdots \\ -\hat{p}_{.1}\hat{p}_{.m} & -\hat{p}_{.2}\hat{p}_{.m} & \cdots & \hat{p}_{.m}(1-\hat{p}_{.m}) \end{bmatrix}, \tag{4.3.1}$$

where $\hat{p}_{.i}$ denotes the estimated probability of finding the i^{th} species. The $(i,k)^{th}$ element of Σ_{yz} is given by

$$\hat{\sigma}_{ik} = \frac{N}{N-1}\left[\sum_{j=1}^{n} z_{jk}\hat{p}_{ji} - \hat{p}_{.i}\sum_{j=1}^{n} z_{jk}\hat{p}_{j.}\right] \tag{4.3.2}$$

and

$$\hat{\Sigma}_{zz} = S_{zz}, \tag{4.3.3}$$

where S_{zz} is the usual sample variance covariance matrix of environmental data, $N = \sum_i \sum_j n_{ij}$, $\hat{p}_{ij} = \frac{n_{ij}}{N}$, $\hat{p}_{i.} = \frac{n_{i.}}{N}$, $n_{i.} = \sum_j n_{ij}$, $\hat{p}_{.j} = \frac{n_{.j}}{N}$ and $n_{.j} = \sum_i n_{ij}$.

In the following we use CCA approach on the Hunter Spider data and compute the species and sites scores. Canonical correlation coefficients resulted by canonical correlation analysis and their contribution towards the variance explained is shown in Table 4.6. It can be seen clearly from the Tables 4.6 and 4.1 that all the canonical

correlation obtained by CCA is similar to what we get by CCPA. The first two canonical correlations capture approximately 88% of the relationship between species and environmental variables. Species and site scores computed by canonical correlation analysis approach are shown in Table 4.7 and 4.8 respectively. Thus, species and site scores calculated by CCA and CCPA do not differ.

Table 4.6: CCA Approach; Hunter Spider Data: Canonical Correlations

Cannonical Correlation	Eigenvalue	Percent	Cumulative %
0.7341518	0.538978865	61.88304447	61.88304447
0.473724	0.224414428	25.76622003	87.64926449
0.2698241	0.072805045	8.359136364	96.00840086
0.1407824	0.019819684	2.275603876	98.28400473
0.1063434	0.011308919	1.298437407	99.58244214
0.0603057	0.003636777	0.417557858	100

Table 4.7: CCA: Species Scores (Standardized Form: Mean = 0, SD = 1)

Species	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6
Ar-I	0.6913	0.9620	-1.0973	-1.5969	0.1113	1.1018
Pa-L	0.6762	-2.5548	1.0037	0.3359	-0.7031	-0.5241
Zo-s	0.6513	0.0298	-0.5823	-1.4449	-1.0598	0.5209
Pa-n	0.5832	0.6254	-0.6738	0.4404	0.6122	-0.4176
Pa-p	0.6035	0.8832	0.0818	0.4278	-0.4371	-1.3998
Au-a	0.5995	0.5546	-0.0967	1.5129	-1.1171	1.1684
Tr-t	0.5220	-0.4104	0.2618	-0.3843	0.2883	-0.2868
Al-c	0.4805	0.0185	0.7858	0.8298	1.8411	1.2281
Pa-m	-0.1415	0.8376	1.6617	-1.0022	0.2425	-0.2734
Al-a	-0.8860	0.3549	1.1386	0.9710	-1.1417	0.1074
Al-f	-1.6397	-0.3144	-0.9652	-0.6755	-0.2649	0.6748
Ar-p	-2.1401	-0.9863	-1.5180	0.5861	1.6282	-1.8998

Table 4.8: CCA: Site Scores (Standardized Form: Mean = 0, SD = 1)

Site	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6
1	1.1158	-0.9066	0.3801	0.0755	0.2837	0.8169
2	0.6289	-1.4222	0.6536	-0.8009	-0.2028	-0.4257
3	0.6153	-1.6232	-0.0219	-1.2911	1.2458	-1.4878
4	0.6153	-1.6232	-0.0219	-1.2911	1.2458	-1.4878
5	1.0496	-0.8916	0.5568	0.3084	0.8261	0.9413
6	0.9283	-1.3203	0.3483	1.0637	0.5093	0.7142
7	0.9764	0.3872	-0.2516	0.4947	-0.9478	0.2506
8	0.7874	-1.4951	0.8415	1.4583	-1.5292	1.1536
9	0.6543	-1.7284	0.1375	-0.2005	0.0344	-0.7245
10	0.7663	0.6338	-1.1704	0.8247	0.2691	-1.0837
11	0.6306	0.5598	-0.4668	-2.0894	-0.6019	2.4838
12	0.9141	1.2013	-0.6685	-1.0749	-1.8326	-0.8595
13	0.3684	0.8630	0.3955	0.4669	-0.6524	-0.3679
14	0.7536	0.7869	-0.9123	0.5245	0.1505	-0.9344
15	0.5872	0.8218	-0.7100	0.7970	1.8040	1.5797
16	0.2360	0.8929	0.7489	0.9329	0.4325	-0.1191
17	0.0765	0.9235	1.3145	0.7876	-0.2402	0.1020
18	-0.1423	0.9592	1.4890	-0.9990	0.1442	-0.9852
19	0.1113	1.2671	1.4681	-0.2854	0.4646	-0.0914
20	-0.6491	0.0651	-2.4634	1.0747	-1.6915	0.5129
21	-0.2561	0.7837	1.7701	0.0068	-0.1368	-0.5180
22	-1.1272	0.9796	-1.1434	-1.2939	0.9515	0.7503
23	-1.4993	0.0314	-0.3465	0.7022	-1.3062	0.7082
24	-1.4878	0.1154	0.3742	0.7932	-0.8917	0.6810
25	-1.8971	-0.0511	-0.4676	-1.9595	1.3386	-1.5877
26	-1.3184	-0.1367	0.2254	1.0187	-1.3249	0.1062
27	-1.3965	0.6655	-0.0657	-0.5929	0.3765	0.7170
28	-2.0416	-0.7388	-1.9935	0.5486	1.2814	-0.8449

IV.4 Population Canonical Correspondence Analysis

The interest here is to provide population versions to the quantities used for canonical correspondence analysis. Consider

Site	y_1	y_2	$\cdots$	y_m	
1	p_{11}	p_{12}	$\cdots$	p_{1m}	$p_{1.}$
2	p_{21}	p_{22}	$\cdots$	p_{2m}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	p_{n1}	p_{n2}	$\cdots$	p_{nm}	$p_{n.}$
	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.m}$	

where p_{ij} is the probability of finding the j^{th} species at the i^{th} site, $p_{.j}$ is the marginal probability of finding the j^{th} species and $p_{i.}$ is the marginal probability of finding species at the i^{th} site.

Following the ideas from Olkin and Tate (1961), we assume for a given site, vector of the environmental variables has multivariate normal distribution:
$\mathbf{z}|(site = k) \sim N(\boldsymbol{\mu_k}, \Sigma)$.

Variance covariance matrix of species and environmental variable $(\mathbf{y}, \mathbf{z})$ is then given by

$$D\begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix} = \begin{bmatrix}\Sigma_{yy} & \Sigma_{yz}\\ \Sigma_{zy} & \Sigma_{zz}\end{bmatrix}, \tag{4.4.1}$$

where

$$\Sigma_{yy} = \begin{bmatrix} p_{.1}(1-p_{.1}) & -p_{.1}p_{.2} & \cdots & -p_{.1}p_{.m} \\ -p_{.1}p_{.2} & p_{.2}(1-p_{.2}) & \cdots & -p_{.2}p_{.m} \\ \vdots & \vdots & \vdots & \vdots \\ -p_{.1}p_{.m} & -p_{.2}p_{.m} & \cdots & p_{.m}(1-p_{.m}) \end{bmatrix} \tag{4.4.2}$$

and δ_{ik}, the covariance of i^{th} species and k^{th} environmental variable, that is, the $(i,k)^{th}$ element of Σ_{yz}, is given by

$$\delta_{ik} = E(y_i z_k) - E(y_i)E(z_k), \tag{4.4.3}$$

where

$$
\begin{aligned}
E(y_i z_k) &= P(y_i \text{ at Site} = 1)E(z_k \text{ at Site} = 1) + \cdots + P(y_i \text{ at Site} = n)E(z_k \text{ at Site } = n) \\
E(y_i z_k) &= p_{1i}\mu_{1k} + p_{2i}\mu_{2k} + \cdots + p_{ni}\mu_{nk} \\
E(y_i z_k) &= \sum_{j=1}^{n} p_{ji}\mu_{jk} \\
E(y_i) &= p_{.i} \\
E(z_k) &= \sum_{j=1}^{n} p_{j.}\mu_{jk}
\end{aligned}
$$

Substituting $E(y_i z_k)$, $E(y_i)$, and $E(z_k)$ in equation 4.4.3 we get

$$\delta_{ik} = \sum_{j=1}^{n} p_{ji}\mu_{jk} - p_{.i}\sum_{j=1}^{n} p_{j.}\mu_{jk}. \tag{4.4.4}$$

It is interesting to note that δ_{ik} is zero when site and species are independent of each other, that is, when $p_{ij} = p_{i.}p_{.j}$ or when all the means of environmental variables are equal, that is, $(\mu_{ij} = \mu_{i'j'} \quad \forall \ i, i', j, j')$. Similarly ψ_{ij}, the $(i,j)^{th}$ element of Σ_{zz}, or the covariance of i^{th} and j^{th} environmental variable, can be computed using

$$
\begin{aligned}
E(z_i z_j) &= \sum_{k=1}^{n} E(z_i z_j | Site = k)P(Site = k) \\
E(z_i z_j) &= \sum_{k=1}^{n} (\sigma_{ij} + \mu_{ki}\mu_{kj})p_{k.} \\
E(z_i z_j) &= \sigma_{ij} + \sum_{k=1}^{n} \mu_{ki}\mu_{kj}p_{k.}
\end{aligned}
$$

$$Cov(z_i, z_j) = \psi_{ij} = \sigma_{ij} + \sum_{k=1}^{n} \mu_{ki}\mu_{kj}p_{k.} - \sum_{k=1}^{n} \mu_{ki}p_{k.} \sum_{k=1}^{n} \mu_{kj}p_{k.} \tag{4.4.5}$$

Let $\sum_{k=1}^{n} \mu_{ki}p_{k.} = \mu_{.i}$ and $\sum_{k=1}^{n} \mu_{kj}p_{k.} = \mu_{.j}$. Then equation 4.4.5 can be written as

$$\psi_{ij} = \sigma_{ij} + \sum_{k=1}^{n} p_{k.}(\mu_{ki} - \mu_{.i})(\mu_{kj} - \mu_{.j}) \tag{4.4.6}$$

If we let $\mathbf{M} = (m_{ij}) = (\mu_{ij} - \mu_{.j}), i = 1, \ldots, n; \ \ j = 1, \ldots, q$ then equation 4.4.2, 4.4.4 and 4.4.6 can be written as

$$
\begin{aligned}
\Sigma_{yy} &= Diag(P_c) - P_c'P_c && (4.4.7) \\
\Sigma_{yz} &= P'\mathbf{M} && (4.4.8) \\
\Sigma_{zz} &= \Sigma + \mathbf{M}'Diag(P_r)\mathbf{M}, && (4.4.9)
\end{aligned}
$$

where $P = (p_{il}), i = 1, \ldots, n, l = 1, \ldots, m$, $P_c = (p_{\cdot 1}, \ldots, p_{\cdot m})$ and $P_r = (p_{1\cdot}, \ldots, p_{n\cdot})$. We note that $\mathbf{1}\Sigma_{yz} = 0$.

Now doing canonical correlation analysis on variance covariance matrices of species and environmental variables $(\mathbf{y}, \mathbf{z})$ will result in canonical variables of species and environmental variables that are highly correlated. The set of, say $\mathbf{v} = (v_1, \ldots, v_m)$ and $\mathbf{w} = (w_1, \ldots, w_q)$, canonical variables are the linearly transformed variables of $\mathbf{y}$ and $\mathbf{z}$. Canonical correlations are the square root of the eigenvalues of $\Sigma_{yy}^{-1/2}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}\Sigma_{yy}^{-1/2}$. Since Σ_{yy} is a singular matrix, some canonical coefficients will be zero and a generalized inverse will be used to compute $\Sigma_{yy}^{-1/2}$. The canonical variables $\mathbf{w}$ is used to compute canonical scores. These canonical scores are called site scores. Species scores are calculated using site scores and is given by equation 4.2.7. All the population parameter are estimated by the corresponding sample counterparts.

We want to make an important remark about the equation 4.4.9 here. In the usual canonical correspondence analysis (CCPA) proposed by Ter Braak, Σ, the variance covariance matrix of environmental variables, is completely ignored due to lack of enough data to estimate it. Hence only the second part of the matrix Σ_{zz}, that is, $\mathbf{M}Diag(P_r)\mathbf{M}'$ is used for the calculation of site and species scores. But in practice, Σ can be estimated using the historical data on the environmental variables and this estimate can be used for performing the above calculations.

IV.4.1 Some Important Special Cases

As described above canonical correlations are the square root of the eigenvalues of $\Sigma_{yy}^{-1/2}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}\Sigma_{yy}^{-1/2}$. Since Σ_{yy} is a singular matrix a generalized inverse could be used to compute $\Sigma_{yy}^{-1/2}$. For the special case, when $m = 2, n = 2$ and $q > 1$ a generalized inverse of Σ_{yy}, can be taken as

$$\Sigma_{yy}^{-} = \begin{bmatrix} \frac{1}{p_{\cdot 1}p_{\cdot 2}} & 0 \\ 0 & 0 \end{bmatrix},$$

and the matrix $B = \Sigma_{yy}^{-1/2}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}\Sigma_{yy}^{-1/2}$ is given by

$$\Sigma_{yy}^{-1/2}\Sigma_{yz} = \begin{bmatrix} \frac{1}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} p_{11} - p_{1\cdot}p_{\cdot 1} & p_{21} - p_{1\cdot}p_{\cdot 2} \\ p_{12} - p_{2\cdot}p_{\cdot 1} & p_{22} - p_{2\cdot}p_{\cdot 2} \end{bmatrix} \begin{bmatrix} \mu_{11} & \cdots & \mu_{1q} \\ \mu_{21} & \cdots & \mu_{2q} \end{bmatrix}$$

$$\Sigma_{yy}^{-1/2}\Sigma_{yz} = \begin{bmatrix} \frac{p_{11}-p_{1\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} & \frac{p_{21}-p_{2\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}$$

$$B = \begin{bmatrix} \frac{p_{11} - p_{1\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} & \frac{p_{21} - p_{2\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix} \Sigma_{zz}^{-1} \begin{bmatrix} \mu^{(1)\prime} & \mu^{(2)\prime} \end{bmatrix}$$

$$\begin{bmatrix} \frac{p_{11} - p_{1\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} & 0 \\ \frac{p_{21} - p_{2\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} \frac{p_{11} - p_{1\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} & \frac{p_{21} - p_{2\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mu^{(1)}\Sigma_{zz}^{-1}\mu^{(1)\prime} & \mu^{(1)}\Sigma_{zz}^{-1}\mu^{(2)\prime} \\ \mu^{(2)}\Sigma_{zz}^{-1}\mu^{(1)\prime} & \mu^{(2)}\Sigma_{zz}^{-1}\mu^{(2)\prime} \end{bmatrix}$$

$$\begin{bmatrix} \frac{p_{11}-p_{1\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} & 0 \\ \frac{p_{21}-p_{2\cdot}p_{\cdot 1}}{\sqrt{p_{\cdot 1}p_{\cdot 2}}} & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} \frac{\mathbf{M}\Sigma_{zz}^{-1}\mathbf{M}'}{p_{\cdot 1}p_{\cdot 2}} & 0 \\ 0 & 0 \end{bmatrix}, \text{ where } \mathbf{M} = [(p_{11} - p_{1\cdot}p_{\cdot 1})\mu^{(1)} + (p_{21} - p_{2\cdot}p_{\cdot 1})\mu^{(2)}]$$

$$B = \begin{bmatrix} \frac{(p_{11} - p_{1\cdot}p_{\cdot 1})^2}{p_{\cdot 1}p_{\cdot 2}}[\mu^{(1)} - \mu^{(2)}]\Sigma_{zz}^{-1}[\mu^{(1)} - \mu^{(2)}]' & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence the eigenvalue of matrix B is given by

$$\lambda = \frac{(p_{11} - p_{1\cdot}p_{\cdot 1})^2}{p_{\cdot 1}p_{\cdot 2}}[\mu^{(1)} - \mu^{(2)}]\Sigma_{zz}^{-1}[\mu^{(1)} - \mu^{(2)}]'.$$

Note that when the two mean vectors are same or/and when the independence holds in the contingency table, i.e. $p_{11} = p_{1\cdot}p_{\cdot 1}$, the eigenvalue $\lambda = 0$. In general for $n > 1$ and the same choice for m and q, matrix B is given by

$$B = \begin{bmatrix} \frac{\mathbf{M}\Sigma_{zz}^{-1}\mathbf{M}'}{p_{\cdot 1}p_{\cdot 2}} & 0 \\ 0 & 0 \end{bmatrix}, \tag{4.4.10}$$

where

$$\mathbf{M} = (p_{11} - p_{1\cdot}p_{\cdot 1})\mu^{(1)} + (p_{21} - p_{2\cdot}p_{\cdot 1})\mu^{(2)} + \ldots + (p_{n1} - p_{n\cdot}p_{\cdot 1})\mu^{(n)}.$$

Note as before that if all the mean vectors are the same or/and when the conditional independence holds at each table then the eigenvalue will be zero. Although we have not pursued here, one can develop tests for testing the eigenvalues to be zero which would in turn test for independence or equality of the means.

IV.5 Canonical Correspondence Analysis of Longitudinal Data

In this section we will discuss how to perform canonical correspondence analysis (CCPA) when we have repeated data. Suppose we have a fixed number n sites, m species vector $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ and q environmental variables vector $\mathbf{z} = (z_1, z_2, \ldots, z_q)$, that were observed at t time periods.

	Time Period = 1				Time Period = k	Time Period = t				
Site	y_1	y_2	$\cdots$	y_m	$\cdots$	y_1	y_2	$\cdots$	y_m	
1	p_{111}	p_{121}	$\cdots$	p_{1m1}	$\cdots$	p_{11t}	p_{12t}	$\cdots$	p_{1mt}	$p_{1..}$
2	p_{211}	p_{221}	$\cdots$	p_{2m1}	$\cdots$	p_{21t}	p_{22t}	$\cdots$	p_{2mt}	$p_{2..}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	p_{n11}	p_{n21}	$\cdots$	p_{nm1}	$\cdots$	p_{n1t}	p_{n2t}	$\cdots$	p_{nmt}	$p_{n..}$
	$p_{.11}$	$p_{.21}$	$\cdots$	$p_{.m1}$	$\cdots$	$p_{.1t}$	$p_{.2t}$	$\cdots$	$p_{.mt}$	1

Here p_{ijk} is the probability of finding j^{th} species at i^{th} site at k^{th} time period, $p_{.jk}$ is the probability of finding j^{th} species at k^{th} time period and $p_{i..}$ is the probability of finding species at i^{th} site. Similarly we have a vector of the environmental variables, $\mathbf{z} = (z_1, z_2, \ldots, z_q)$ observed over t time periods i.e. $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_t)$. Assuming we have a repeated effect in both variables i.e. in both species abundance and environmental variables, we can proceed as follows.

In this case the variance covariance matrix of species and environmental variables $(\mathbf{Y}, \mathbf{Z})$ is given by

$$\Sigma_{(m+q)t\times(m+q)t} = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{bmatrix}, \tag{4.5.1}$$

where

$$\Sigma_{zz} = \Omega_{t\times t} \otimes \Sigma + \mathbf{M}' Diag(P_r)\mathbf{M} \tag{4.5.2}$$

$$\hat{\Sigma}_{yy} = \begin{bmatrix} \hat{p}_{.11}(1-\hat{p}_{.11}) & -\hat{p}_{.11}\hat{p}_{.21} & \cdots & -\hat{p}_{.11}\hat{p}_{.mt} \\ -\hat{p}_{.11}\hat{p}_{.21} & \hat{p}_{.21}(1-\hat{p}_{.21}) & \cdots & -\hat{p}_{.21}\hat{p}_{.mt} \\ \vdots & \vdots & \vdots & \vdots \\ -\hat{p}_{.11}\hat{p}_{.mt} & -\hat{p}_{.21}\hat{p}_{.mt} & \cdots & \hat{p}_{.mt}(1-\hat{p}_{.mt}) \end{bmatrix} \tag{4.5.3}$$

The covariance of l^{th} species at k^{th} time period or overall mt species categories and r^{th} environmental variable at k^{th} time period or overall qt environmental categories is given by $(l,r)^{th}$ element of Σ_{yz}.

$$\hat{\sigma_{lr}} = \sum_{j=1}^{n} z_{jrk}\hat{p}_{jlk} - \hat{p}_{.lk}\sum_{j=1}^{n} z_{jrk}\hat{p}_{j..} \quad l = 1,\ldots,mt;\ r = 1,\ldots,qt. \tag{4.5.4}$$

The estimates of $\Omega_{t\times t}\otimes\Sigma$ can be obtained from the environmental data observed over t time periods by maximizing the log likelihood of multivariate normal function. See Naik and Rao (2001). We can then calculate the species and site scores as suggested in IV.3. After fitting the above variance-covariance relationship and calculating the species scores we can plot the species scores for all the time periods. If these profiles are homogeneous then we can take summary statistics of species scores and call it final species scores.

In the following example, we use simulated data to to illustrate the methods discussed in this section.

IV.6 An Example: Analysis of Simulated Data

A (6×4), site by species, simulated contingency table is shown in Table 4.12. A simulated set of correlated multivariate normal data considered as data on environmental variables for 10 time periods is given in Table 4.13. To test the time effect on contingency table we can use the method as discussed in section III.4. The output of correlated linear model on $\boldsymbol{\lambda}_1$ is shown in Table 3.15. The small p-value $(p - val = 0.0020)$ suggest that there is a repeated effect in the contingency table. To test the repeated effect in simulated environmental data we can test

$$H_o : \Sigma^* = I_{tt} \otimes \Sigma \text{ Vs } H_o : \Sigma^* = \Omega_{tt} \otimes \Sigma.$$

The likelihood ratio test (LRT) as discussed in II.4 can be used to test above hypothesis. The result from LRT test statistics is shown in Table 4.11. The likelihood ratio test statistic is 30.320489 and we compare this with 1 degree of freedom chi-square. The small p-value $(3.6624E-8 \approx 0)$ concludes that there is a repeated effect in the environmental variables.

Hence we have to calculate the species and site scores as described in section IV.5. In this example because of longitudinal study of environmental variables we

have sufficient data to estimate the variance covariance structure of environmental variables. The resulted canonical correlation between species and environmental variables are shown in Table 4.9. The maximum and minimum canonical correlations are 0.130619 and 0.000071 respectively. The first two canonical variables capture approximately 98% of the relationship between species and environmental variables. Species and site scores after fitting the structure is shown in Tables 4.14 and 4.10. Species score profiles after fitting of the variance covariance structure is shown in Figure 4.2. These profiles can be used to get a better understanding of species scores at different time periods. As in this case Figure 4.2 suggests that the ranking of species 1 is higher as compared to other species for all the time periods. It can also be inferred from the species profile that score of 3^{rd} specie is almost constant for all time periods.

IV.7 Concluding Remarks

In this chapter we have considered canonical correspondence analysis (CCPA) where the relationship between a set of qualitative and another set of quantitative variables is studied. After providing an introduction of the CCPA we show that this analysis can be performed using canonical correlation analysis (CCA) of a certain matrices. Next using this equivalence relation between CCPA and CCA, we provide a theoretical basis to CCPA which did not exist in the literature. Finally, we provide CCPA method for analyzing repeatedly observed data.

Table 4.9: Simulated Data Example: Canonical Correlations

Cannonical Correlation	Eigen Value	Percent	cum %
0.130619	0.017061	91.01627101	91.01627101
0.035188	0.001238	6.604427847	97.62069885
0.015259	0.000233	1.242998133	98.86369699
0.013306	0.000177	0.9442518	99.80794879
0.00599	0.000036	0.192051214	100
0.000071	0	0	100

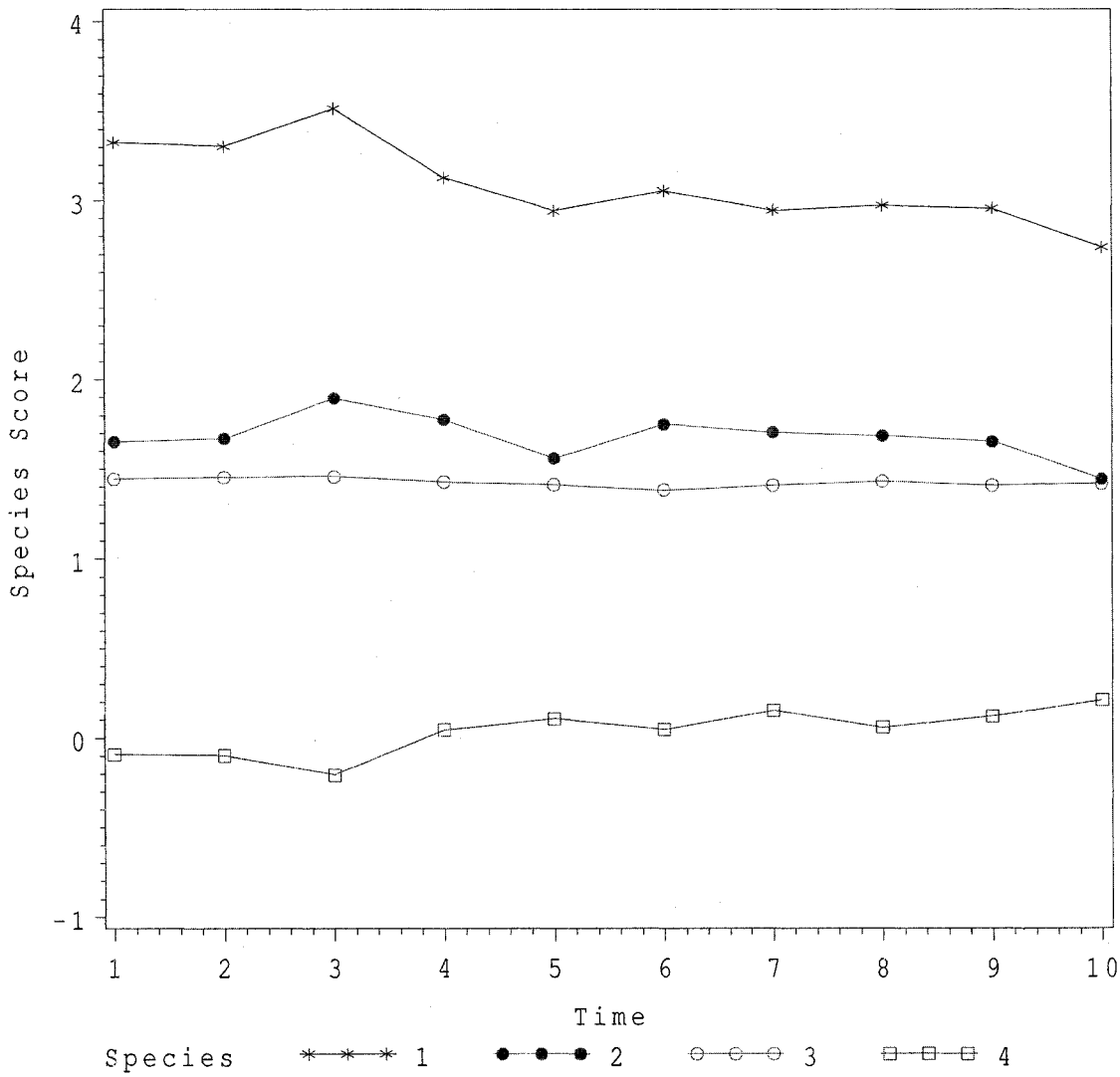

Figure 4.2: Species Scores Profile After Fitting the Variance Covariance Structure.

Table 4.10: Simulated Data Example: Site Scores

Site	Dim1	Dim2
1	1.2198324	-0.561291
2	1.2199394	-0.479067
3	0.5770421	-0.501926
4	0.1992913	1.3983152
5	-0.764373	0.5793593
6	-1.466163	-1.28597

Table 4.11: Hypothesis Testing: Simulated Data

Hypothesis	**Chi Square Test Statistics**	**Dof**	**p-value**
H_o	30.320489	1	$3.6624E-8 \approx 0$

Table 4.12: Simulated Contingency table for 10 Time Period

Site	Time Period = 1				Time Period = 2				Time Period = 3			
	$Y1$	$Y2$	$Y3$	$Y4$	$Y1$	$Y2$	$Y3$	$Y4$	$Y1$	$Y2$	$Y3$	$Y4$
1	69	93	63	50	69	94	62	49	53	85	48	37
2	62	93	59	44	61	94	58	44	46	85	45	31
3	62	104	70	65	61	105	70	65	46	97	55	49
4	77	140	83	93	77	135	82	94	60	121	64	85
5	40	96	59	84	40	97	58	83	28	89	45	66
6	24	76	59	76	24	76	58	76	16	59	45	59
Site	Time Period = 4				Time Period = 5				Time Period = 6			
	$Y1$	$Y2$	$Y3$	$Y4$	$Y1$	$Y2$	$Y3$	$Y4$	$Y1$	$Y2$	$Y3$	$Y4$
1	75	109	69	56	84	108	79	66	74	105	67	56
2	68	109	65	51	78	108	75	59	66	105	65	49
3	68	119	76	71	78	118	87	81	66	117	75	69
4	85	145	89	109	94	151	98	108	82	148	86	105
5	47	111	65	90	56	111	75	101	45	108	65	88
6	29	83	65	83	37	93	75	93	31	81	65	81
	Time Period = 7				Time Period = 8				Time Period = 9			
	$Y1$	$Y2$	$Y3$	$Y4$	$Y1$	$Y2$	$Y3$	$Y4$	$Y1$	$Y2$	$Y3$	$Y4$
1	80	110	74	61	74	100	67	55	79	103	71	60
2	72	110	70	57	65	100	63	50	71	103	68	54
3	72	122	80	75	65	112	74	68	71	114	80	73
4	89	168	94	110	81	164	85	100	86	159	92	103
5	50	113	70	94	46	103	63	87	50	106	68	93
6	36	87	70	87	32	80	63	80	35	84	68	84
	Time Period = 10											
	$Y1$	$Y2$	$Y3$	$Y4$								
1	94	110	88	75								
2	87	110	83	68								
3	87	121	96	90								
4	103	166	107	110								
5	65	112	83	109								
6	47	102	83	102								

Table 4.13: Simulated Environmental Variables for 10 Time Period

Site	Time Period = 1				Time Period = 2			
	$Z1$	$Z2$	$Z3$	$Z4$	$Z1$	$Z2$	$Z3$	$Z4$
1	-0.333	-0.209	0.854	0.050	-0.568	-0.662	0.170	1.584
2	-1.263	-0.170	0.650	0.789	-1.247	0.724	0.403	0.358
3	0.186	-0.190	-0.477	0.470	-1.239	0.842	0.434	0.314
4	-0.804	-1.940	-1.168	-0.323	0.436	-1.431	-0.238	0.169
5	0.049	0.375	0.666	1.520	0.080	-0.462	-0.399	-0.244
6	-0.986	1.256	1.080	0.856	0.276	1.265	1.180	0.241
Site	Time Period = 3				Time Period = 4			
	$Z1$	$Z2$	$Z3$	$Z4$	$Z1$	$Z2$	$Z3$	$Z4$
1	0.313	-0.986	-0.861	-0.450	0.342	0.779	0.154	0.912
2	-1.860	0.123	-0.173	-0.644	-1.229	0.346	0.744	0.629
3	0.156	0.695	-0.531	0.311	-1.296	0.294	-0.742	1.200
4	0.106	-0.896	0.235	-1.537	-0.127	-0.261	0.148	-1.022
5	0.182	-0.437	-1.296	0.116	1.416	0.139	-0.106	1.061
6	-0.095	-0.669	0.632	-0.857	-0.123	-1.573	0.318	-0.834
Site	Time Period = 5				Time Period = 6			
	$Z1$	$Z2$	$Z3$	$Z4$	$Z1$	$Z2$	$Z3$	$Z4$
1	-1.397	-1.371	-1.506	-0.850	-0.428	-1.580	-1.762	-0.950
2	-1.904	-0.408	-0.182	0.935	1.439	1.143	0.948	-0.054
3	-1.058	-0.753	-0.153	0.669	-0.693	0.096	-0.028	0.306
4	0.344	0.783	0.592	1.097	0.745	-0.307	0.007	0.771
5	1.382	-0.396	0.451	-0.157	0.440	-1.008	0.217	-0.678
6	-1.028	0.098	1.421	-0.643	-2.240	-2.199	0.499	0.591
Site	Time Period = 7				Time Period = 8			
	$Z1$	$Z2$	$Z3$	$Z4$	$Z1$	$Z2$	$Z3$	$Z4$
1	1.069	0.109	-0.343	2.220	1.158	-0.437	0.182	-0.143
2	-0.186	-0.085	1.648	0.869	-0.137	-0.424	0.509	0.938
3	-0.480	-0.656	0.117	0.802	-0.800	0.120	0.515	-0.041
4	1.508	1.236	0.487	0.932	0.839	1.042	-0.104	-0.103
5	0.437	0.014	-1.100	0.010	0.738	-0.588	-0.718	-0.637
6	-0.132	1.197	1.313	1.759	0.880	0.631	1.024	1.283
Site	Time Period = 9				Time Period = 10			
	$Z1$	$Z2$	$Z3$	$Z4$	$Z1$	$Z2$	$Z3$	$Z4$
1	0.668	0.405	0.171	0.062	-0.019	1.795	2.061	-0.449
2	0.037	0.143	1.233	0.386	0.577	0.834	0.963	0.116
3	-0.862	-0.390	-0.285	0.297	-0.291	0.047	1.186	0.593
4	-0.269	0.082	0.372	-0.186	1.572	0.678	1.592	0.205
5	0.134	-0.291	-0.812	-1.961	0.679	0.273	0.353	0.086
6	1.087	0.098	-0.778	0.460	-0.179	-1.424	-1.414	-0.719

Table 4.14: Simulated Data Example: Species Scores

Site	Dim1	Dim2
Y1	3.327406	0.0366452
Y2	1.6533054	0.22165
Y3	1.4446456	-1.763773
Y4	-0.091516	-0.054695
Y5	3.3060126	0.1208377
Y6	1.6720557	-0.154116
Y7	1.4552013	-1.760965
Y8	-0.09651	0.0406095
Y9	3.5176559	0.5331461
Y10	1.8987538	0.4119444
Y11	1.4593547	-1.732534
Y12	-0.205689	1.8258446
Y13	3.1318553	-0.000828
Y14	1.777647	-0.538074
Y15	1.4289463	-1.95384
Y16	0.045522	0.3905538
Y17	2.945997	-0.489913
Y18	1.5609583	-0.648671
Y19	1.4138317	-2.318864
Y20	0.1084241	-0.896651
Y21	3.0573451	-0.422625
Y22	1.7510877	-0.111497
Y23	1.3837949	-2.154246
Y24	0.0479367	0.2603134
Y25	2.9492438	-0.599125
Y26	1.707198	0.5135694
Y27	1.4111945	-2.062872
Y28	0.154055	-0.164113
Y29	2.976622	-0.51178
Y30	1.6876132	1.1155571
Y31	1.4334355	-2.09477
Y32	0.0602423	-0.103159
Y33	2.9599476	-0.708904
Y34	1.654596	0.4819819
Y35	1.4110358	-2.033619
Y36	0.1225163	-0.38352
Y37	2.7411247	-0.992219
Y38	1.4453846	-0.367862
Y39	1.4203872	-2.440335
Y40	0.2132843	-1.746722

CHAPTER V

CA FOR HIGHER DIMENSIONS

V.1 Introduction

Although canonical correlation and other methods discussed in the previous chapters are general methods, their utility when dealing with really large data sets have not been well studied. In this era of internet, genomics, and proteomics, information available in the form of data are explosive in nature. Hence it is important that we look at analysis of at least some of such large data sets using some of the methods we have studied thus far.

In this chapter, we work with a high dimensional data set in the field of language processing. The data are in the form of contingency tables and usually very sparse in nature. We will use correspondence analysis to analyze these data and compare its performance with a well established method in this are named latent semantic analysis (LSA).

In the next section, we will review latent semantic analysis (LSA) which is a popular method of analysis of the data among the practitioners in natural language processing. In Section 3 we show that a correspondence analysis can be used for this purpose. We compare the two methods in Section 4 and provide some guidelines on which method is better in what situation.

V.2 Latent Semantic Analysis

Latent semantic analysis (LSA) is a technique in natural language processing, in particular in vectorial semantics, invented by Deerwester, Dumais, Furnas, Landauer and Harshman (1990). LSA analyze relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. It is a fully automatic mathematical/statistical technique for extracting and inferring relation of expected contextual usage of words in passages of discourse. It is not a traditional language processing or artificial programme. It uses no humanly constructed dictionaries. It takes as its input only the raw text parsed into words

defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs (Landauer, Foltz and Laham, 1998). The underlying idea is that the totality of information about all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and set of words to each other. In one sense it can be said that LSA represents the meaning of a word as a kind of average of the meanings of all the passages in which it appears, and the meaning of a passage as a kind of average of the meaning of all the words it contains.

In LSA the first step is to represent the text as a matrix, called term-document matrix, in which each row stands for a unique word and each column stands for a text passage or other context. Each cell of this matrix contains the frequency with which a word in its row appears in the passage denoted by its column. If we represent such a matrix by **A** then

$$\mathbf{A} = [a_{ij}],$$

where a_{ij} denotes the frequency in which the i^{th} term occurs in the j^{th} document. Since every word does not normally appear in each document, the matrix **A** is usually quite sparse. Next, the cell entries are subjected to a preliminary transformation in which each cell frequency is weighted by a function that expresses both the word's importance in the particular passage and the degree to which the word type carries information in the domain of discourse in general. Thus, a_{ij} is represented as

$$a_{ij} = L(i,j)G(i),$$

where $L(i,j)$ is called local weighting for term i in document j, and $G(i)$ is called global weighting for term i. Some popular local and global weighting schemes are given in Tables 5.1 and 5.2 respectively.

where

$$\delta(a_{ij}) = \begin{cases} 1 & \text{if } a_{ij} > 0 \\ 0 & \text{if } a_{ij} = 0 \end{cases}$$

p_{ij} = probability of the i^{th} term in the j^{th} document and ndocs = Total number of documents in the collection. It has been seen in practice that LogEntropy weighting scheme,

$$a_{ij} = log(a_{ij} + 1) \times \left[1 - \left(\sum_j \frac{p_{ij} log(p_{ij})}{log(\text{ndocs})}\right)\right]$$

Table 5.1: Formulas for Local Term Weights

Symbol	Name	Formula
b	Binary, Salton and McGill (1983)	$\delta(a_{ij})$
l	Logarithmic, Harman (1992)	$log(1+a_{ij})$
n	Aug. normalized term frq, Salton and McGill (1983); Harman (1992)	$\frac{\left[\delta(a_{ij}) + (\frac{a_{ij}}{\max\limits_{k} a_{kj}})\right]}{2}$
t	Term Frequency, Salton and McGill (1983)	a_{ij}

is the best. Next, LSA involves determining the singular value decomposition (SVD) of the matrix $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}',$$

where it is well known that any rectangular matrix can be so decomposed perfectly, using no more factors than the smallest dimension of the original matrix. When fewer than the necessary number of factors are used, the reconstructed matrix is a least-squares best fit. One can reduce the dimensionality of the solution simply by deleting coefficients in the diagonal matrix, ordinarily starting with the smallest. Thus k-dimensional best fit of $\mathbf{A}$ would be

$$A_k = U_k \Sigma_k V_k'$$

The truncated SVD captures most of the important underlying structure in the association of terms and documents, yet at the same time removes the noise or variability in word usage. The result of the SVD is a k-dimensional vector space containing a vector for each term and each document. Finally term and document vectors are plotted in $k-$dimensional space on the same graph.

One can interpret the analysis performed by SVD geometrically. The location of term vectors reflects the correlations in their usage across documents. Similarly, the location of document vectors reflects correlations in the terms used in the documents. In this space the cosine or dot product between vectors corresponds to their

Table 5.2: Formulas for Global Term Weights

Symbol	Name	Formula
e	Entropy, Dumais (1991)	$1 - \left(\sum_j \frac{p_{ij} log(p_{ij})}{log(\text{ndocs})}\right)$
f	Inverse document frequency (IDF), Dumais (1991); Salton and McGill (1983)	$log_2\left(\frac{\text{ndocs}}{\sum_j \delta(a_{ij})}\right) + 1$
g	GfIdf, Dumais (1991)	$\frac{\sum_j a_{ij}}{\sum_j \delta(a_{ij})}$
n	Normal, Dumais (1991)	$\sqrt{\frac{1}{\sum_j a_{ij}^2}}$
p	Probabilistic Inverse, Salton and McGill (1983); Harman (1992)	$log\left(\frac{\text{ndocs}-\sum_j \delta(a_{ij})}{\sum_j \delta(a_{ij})}\right)$

estimated semantic similarity. Thus, by determining the vectors of two pieces of textual information, we can determine the semantic similarity between them.

V.3 Illustration of LSA

In this section we will explain LSA technique through examples.

V.3.1 Example 1

This example uses 17 book titles from book reviews published in the *SIAM Review,* Volume 54. All the underlined words in Table 5.3 denote keywords used as referents to the book titles. The parsing rule used for this example required that keywords appear in more than one book title.

The term-document matrix $\mathbf{A}_{16\times17}$, corresponding to text in Table 5.3, is shown in Table 5.4. The elements of this matrix are the frequencies in which a term occurs in a document or book title. For example, in book title B1, the first column of the term-document matrix $\mathbf{A}_{16\times17}$, the terms *equations* and *integral* occur once. Transforming

Table 5.3: Database of Titles from Books Received in SIAM Review

Label	Title
B1	A course on Integral Equations
B2	Attractors for semigroups and Evolution Equations
B3	Automatic Differentiation of Algorithm: Theory, Implementation, and Application
B4	Geometrical Aspects of Partial Differential Equations
B5	Ideals, Varieties, and Algorithms - An Introduction to Computational Algebraic Geometry and Commutative Algebra
B6	Introduction to Hamiltonian Dynamical Systems and the $N-$Body Problem
B7	Knapsack Problems: Algorithms and Computer Implementations
B8	Methods of Solving Singular Systems of Ordinary Differential Equations
B9	Nonlinear Systems
B10	Ordinary Differential Equations
B11	Oscillation Theory for Neutral Differential Equations with Delay
B12	Oscillation Theory of Delay Differential Equations
B13	Pseudodifferential Operators and Nonlinear Partial Differential Equations
B14	Sinc Methods for Quadrature and Differential Equations
B15	Stability of Stochastic Differential Equations with Respect to Semi-Martingales
B16	The Boundary Integral Approach to Static and Dynamic Contact Problems
B17	The Double Mellin-Barnes Type Integrals and Their Applications to Convolution Theory

the elements of term-document matrix $\mathbf{A}_{16\times17}$ according to LogEntropy weighting scheme we get transformed matrix $\mathbf{A}_{16\times17}$, truncated to one decimal place, as shown in Table 5.5. Next, SVD decomposition of transformed matrix $\mathbf{A}_{16\times17}$ gives

$$\mathbf{A}_{16\times17} = U_{16\times k}\Sigma_{k\times k}V'_{k\times17}$$

Now choosing $k = 2$, the truncated SVD of the transformed matrix $\mathbf{A}_{16\times17}$ will give rank-2 approximation $\mathbf{A}_2$

$$\mathbf{A} \approx \mathbf{A}_2 = U_2\Sigma_2V_2',$$

where $U_2 =$

$$\begin{bmatrix} 0.0072051 & 0.4077252 \\ 0.0121796 & 0.3464152 \\ 0.1219809 & 0.1798224 \\ 0.6134575 & -0.09226 \\ 0.7228448 & -0.10095 \\ 0.0061875 & 0.3073185 \\ 0.0405602 & 0.237913 \\ 0.0025155 & 0.0890314 \\ 0.1099096 & -0.093971 \\ 0.0574432 & -0.054779 \\ 0.1099096 & -0.093971 \\ 0.1219809 & 0.1798224 \\ 0.1063124 & -0.082826 \\ 0.0045159 & 0.2143652 \\ 0.0704532 & -0.037973 \\ 0.1605313 & 0.6296756 \end{bmatrix}$$

$V_2 =$

$$\begin{bmatrix} 0.1771327 & 0.0387424 \\ 0.1698369 & -0.048172 \\ 0.0352476 & 0.6100063 \\ 0.324861 & -0.117421 \\ 0.0017018 & 0.1781205 \\ 0.013891 & 0.0936102 \\ 0.0031065 & 0.3279531 \\ 0.3558455 & -0.165735 \\ 0.0219399 & -0.03182 \\ 0.3254413 & -0.121073 \\ 0.378056 & 0.2744157 \\ 0.378056 & 0.2744157 \\ 0.3341281 & -0.135369 \\ 0.3254413 & -0.121073 \\ 0.30771 & -0.090284 \\ 0.0081081 & 0.1652258 \\ 0.0402492 & 0.447279 \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} 5.3477028 & 0 \\ 0 & 2.633105 \end{bmatrix}.$$

Using the first column of U_2 multiplied by first singular value, σ_1, for the $x-$coordinates and the second column of U_2 multiplied by second singular value, σ_2, for the $y-$coordinates, the terms can be represented on the Cartesian plane. Similarly, the first column of V_2 scaled by σ_1 are the $x-$coordinates and the second column of V_2 scaled by σ_2 are the $y-$coordinates for the documents. Figure 5.1 is a two-dimensional plot of the term-document matrix $\mathbf{A}_{16\times17}$.

Table 5.4: 16×17 *Term-Document Matrix*

Terms	Documents																
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17
Algorithms	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
Application	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Delay	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Differential	0	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
equations	1	1	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
implementation	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
integral	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
introduction	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
methods	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
nonlinear	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ordinary	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
oscillation	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
partial	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
problem	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
systems	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0
theory	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1

Table 5.5: LogEntropy Weighting Scheme $A_{16\times17}$ *Term-Document Matrix*

Terms	Documents																
Algo	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
App	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9
Delay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.9	0.0	0.0	0.0	0.0	0.0
Diff	0.0	0.0	0.0	1.2	0.0	0.0	0.0	1.2	0.0	1.2	1.2	1.2	1.2	1.2	1.2	0.0	0.0
eq	1.3	1.3	0.0	1.3	0.0	0.0	0.0	1.3	0.0	1.3	1.3	1.3	1.3	1.3	1.3	0.0	0.0
impl	0.0	0.0	0.9	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
int	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
intro	0.0	0.0	0.0	0.0	0.9	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metd	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0
non	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0
ord	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
osc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.9	0.0	0.0	0.0	0.0	0.0
par	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0
prob	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
sys	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
theory	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0

Since the initial data are in the form of a contingency table, we can adopt correspondence analysis to get a two-dimensional representation of terms and documents. The correspondence matrix $P(\frac{n_{ij}}{N})$ can be calculated easily. The coordinates for the 16 term profiles and 17 document profiles in 2-dimensional space are given by $\mathbf{F}_{16\times 2}$ and $\mathbf{G}_{17\times 2}$ respectively.

$$\mathbf{F}_{16\times 2} = \begin{bmatrix} 1.7175 & 0.106 \\ 1.089 & -1.3636 \\ -0.5192 & -0.9361 \\ -0.7661 & -0.0243 \\ -0.696 & -0.1051 \\ 1.6195 & -0.4677 \\ 0.6694 & -0.839 \\ 1.5487 & 1.5817 \\ -0.7886 & 0.4249 \\ -0.5829 & 1.4899 \\ -0.7886 & 0.4249 \\ -0.5192 & -0.9361 \\ -0.8887 & 0.3891 \\ 1.4229 & 0.6288 \\ 0.0597 & 1.6479 \\ 0.2849 & -1.1499 \end{bmatrix} \qquad \mathbf{G}_{17\times 2} = \begin{bmatrix} -0.0144 & -0.5753 \\ -0.7533 & -0.1281 \\ 1.2748 & -0.876 \\ -0.8482 & 0.1055 \\ 1.7677 & 1.0284 \\ 1.0938 & 1.5674 \\ 1.7174 & 0.1085 \\ -0.645 & 0.5772 \\ -0.2832 & 1.912 \\ -0.8121 & 0.12 \\ -0.4796 & -0.7681 \\ -0.4796 & -0.7681 \\ -0.7939 & 0.5331 \\ -0.8121 & 0.12 \\ -0.7913 & -0.0788 \\ 1.1324 & -0.1281 \\ 0.7373 & -1.3619 \end{bmatrix}$$

Figure 5.2 is a two-dimensional plot generated by correspondence analysis of the term-document matrix $\mathbf{A}_{16\times 17}$.

Note that in Figure 5.1 documents (B10, B14, B15, B13) and terms pertaining to *differential equations* are clustered around the $x-$axis. Similarly we can see that B11 and B12 form a cluster. Such grouping suggest that subset of book titles contains titles similar in meaning. Similar things can be seen in Figure 5.2 as well.

V.3.2 Example 2

In Chapter 3 we have shown correspondence analysis of Parents' socioeconomic status on children's mental health. See details in III.2.1. Latent semantic analysis (LSA) can also be used to analyze these data. In the following we will provide a two dimensional representation of parent's socioeconomic status and children's mental impairment using LSA. First transforming the data given in Table 3.1 according to LogEntropy weighting scheme we get

$$\mathbf{A}_{6\times 4} = \begin{bmatrix} 8.2434487 & 8.9928529 & 8.0521928 & 7.6031504 \\ 7.9736259 & 8.9425983 & 7.8693324 & 7.292464 \\ 8.0217832 & 9.2130581 & 8.277053 & 8.121414 \\ 8.463827 & 9.7764036 & 8.5945178 & 8.9834729 \\ 7.0574758 & 8.9612387 & 7.8322626 & 8.5400093 \\ 5.9968356 & 8.2970273 & 7.7745028 & 8.2970273 \end{bmatrix}.$$

Next, truncated SVD decomposition of transformed matrix $\mathbf{A}_{6\times 4}$ in two dimension gives

$$\mathbf{A} \approx \mathbf{A}_2 = U_2 \Sigma_2 V_2'$$

where $U_2 = \begin{bmatrix} 0.4075903 & -0.417695 \\ 0.3977873 & -0.425438 \\ 0.417101 & -0.146411 \\ 0.444232 & -0.020594 \\ 0.4023228 & 0.3591943 \\ 0.377434 & 0.7026044 \end{bmatrix}$ $V_2 = \begin{bmatrix} 0.4641468 & -0.752015 \\ 0.5484784 & -0.032818 \\ 0.4896015 & 0.0919584 \\ 0.4939936 & 0.6518749 \end{bmatrix}$

and

$$\Sigma_2 = \begin{bmatrix} 40.375905 & 0 \\ 0 & 1.9094061 \end{bmatrix}.$$

The co-ordinates of parent's socioeconomic status and children's mental impairment in two-dimensional space is shown in Table 5.6. Figure 5.3 is a two-dimensional plot generated by latent semantic analysis (LSA) of the socioeconomic status by mental health of children data.

It is interesting to note that categories are ordered in socioeconomic status by mental health of children data and the order is maintained in both two-dimensional representation, i.e, in Figure 3.1 and 5.3. The point corresponding to status category 6 (low status) is closest to IMPAIRED, followed by other categories in a decreasing

Table 5.6: LSA: 2-Dimensional Coordinates of Socioeconomic Status by Mental Health of Children Data

Label	Dim1	Dim2
1	16.4568	-0.79755
2	16.061	-0.81233
3	16.8408	-0.27956
4	17.9363	-0.03932
5	16.2441	0.68585
6	15.2392	1.34156
WELL	18.7403	-1.4359
MILD	22.1453	-0.06266
MODERATE	19.7681	0.17559
IMPAIRED	19.9454	1.24469

order. Similarly, point corresponding to the status 1 (High) is closest to the point corresponding to WELL, followed by other status categories in the increasing order. The points representing the status categories 1 and 2 form a cluster in both two-dimensional representation. Hence these categories may be clubbed together to form one group. The categories corresponding to the mental status of children also follow an order from IMPAIRED to WELL. The two middle categories are quite close to each other, but there is a clear distinction between the other categories. This is captured by more in latent semantic analysis (LSA) representation than by correspondence analysis representation.

V.4 Correspondence Analysis of High Dimension Data

In section V.3 we have seen that data given in the form of contingency table can be analyzed by both correspondence analysis (CA) and latent semantic analysis (LSA). We can get two types of graphical representation of the same data and it is generally difficult to decide which one is better. Interpretation of graphical representation of the given data in the form of contingency table is very subjective. Taking the data from information retrieval area we will provide ceratin analysis to decide on which of the two representation is better.

There is now a huge amount of information stored in electronic format. This

includes books, newspapers, magazines, academic journals, web sites and on-line databases. The World Wide Web (WWW) has made this electronic information accessible to a large number of people. The purpose of an Information Retrieval (IR) system is to help people find relevant information when they request it. The objective of correspondence analysis (CA) and latent semantic analysis (LSA) is to represent the relationship between the categories, in this case various terms (t_i) in different documents (d_j), of two variables in lower dimensional space. If we can retrieve more relevant documents for a given query in the low dimensional space using, say correspondence analysis (CA) then we can say it is better than latent semantic analysis (LSA).

For the purposes of information retrieval, a query must be represented as a vector in low dimensional space and compared to documents. A query or *pseudo-document* is a set of words. Let $\mathbf{X}$ be term-document matrix. Then according to latent semantic analysis (LSA), $k-$dimensional space is given by SVD decomposition of matrix $\mathbf{X}$ and taking $k-$largest singular triplets.

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{T}_{t\times k}\mathbf{\Sigma}_{k\times k}\mathbf{D}_{k\times d} \tag{5.4.1}$$

Here $\mathbf{T}_{t\times k}$: term coordinates in $k-$dimensional space and $\mathbf{D}_{k\times d}$: document coordinates in $k-$dimensional space. Query can be represented by equation 5.4.2 as given by Deerwester, Dumais, Furnas, Landauer and Harshman (1990).

$$\hat{\mathbf{D}}_q = \mathbf{q}'\mathbf{T}_{t\times k}\mathbf{\Sigma}^{-1}_{k\times k}, \tag{5.4.2}$$

where q is simply the vector of words in the query. Thus, the query vector is a weighted sum of its constituent term vector. The query vector can then be compared to all existing document vectors, and documents ranked by their similarity to the query. One measure of similarity is the cosine between the query vector and document vector. For example, suppose we are interested in the documents that pertain to *application theory* in Example 1 of section V.3. Query representation in 2-dimensional

space is given by equation 5.4.2

$$\hat{\mathbf{D}}_q = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}' \begin{bmatrix} 0.0072051 & 0.4077252 \\ 0.0121796 & 0.3464152 \\ 0.1219809 & 0.1798224 \\ 0.6134575 & -0.09226 \\ 0.7228448 & -0.10095 \\ 0.0061875 & 0.3073185 \\ 0.0405602 & 0.237913 \\ 0.0025155 & 0.0890314 \\ 0.1099096 & -0.093971 \\ 0.0574432 & -0.054779 \\ 0.1099096 & -0.093971 \\ 0.1219809 & 0.1798224 \\ 0.1063124 & -0.082826 \\ 0.0045159 & 0.2143652 \\ 0.0704532 & -0.037973 \\ 0.1605313 & 0.6296756 \end{bmatrix} \begin{bmatrix} 5.3477028 & 0 \\ 0 & 2.633105 \end{bmatrix}^{-1}$$

$$\hat{\mathbf{D}}_q = [0.0323, 0.3707].$$

The two-dimensional representation of query vector is shown in Figure 5.4. This query vector is then compared to all the documents in the database and ranked based on their cosine. This is shown in Table V.4.

Similarly in correspondence analysis query or *pseudo-document* can be represented by equation 5.4.3

$$\hat{\mathbf{D}}_q = \mathbf{q}'\mathbf{D}_r^{-1}\mathbf{A}_{t\times k}\mathbf{\Lambda}_{k\times k}^{-1}. \tag{5.4.3}$$

Query coordinates in two-dimensional space is given by

$$\hat{\mathbf{D}}_q = [34.6565, \quad -61.42546]$$

and ranking of documents based on their cosine is shown in Table V.4.

The performance of information-retrieval, as discussed by Berry and Browne (2005), is often summarized in terms of two parameters: precision and recall.
Recall is the proportion of all relevant documents in the collection that are retrieved

by the system, that is,

$$R = \frac{Doc_r}{N_r}, \tag{5.4.4}$$

where Doc_r is the number of relevant documents retrieved and N_r is the total number of relevant documents in the collection. *Precision* is the proportion of relevant documents in the set return to the user, that is,

$$P = \frac{Doc_r}{Doc_t}, \tag{5.4.5}$$

where Doc_t is the total number of documents retrieved. Precision is calculated for several levels of recall, and averaged over queries. For our purpose of comparison between latent semantic analysis (LSA) and correspondence analysis (CA) we have taken *MED* collection. The Medline (also referred to as *MED*, *MEDLARS* or *MED1033*) was the commonly studied collection of medical abstracts. It consists of 1033 documents and 30 queries and frequently used in the IR literature. Some characteristics of the *MED* dataset are shown in Table V.4. The number of unique terms can vary somewhat because different term-processing algorithms were used in the different systems. In our case we have counted only those terms which occur in more than one document and not on SMART's stop list of common words. Stop lists are lists of words that have little or no value as a search item. SMART's stop list is a list of word developed by SMART system at Cornell University (see ftp://ftp.cs.cornell.edu/pub/smart/english.stop).

Each cell of Table 5.10 shows average precision of correspondence analysis (CA) method, average taken over all 30 queries, for a given Doc_t. The last column of Table 5.10 represent average precision (Avp_{qd}), averaged over all 30 queries and 9 level of Doc_t. Similarly Table 5.11 shows average precision of latent semantic analysis (LSA) method. Figure 5.5 shows average precision as a function of dimension for latent semantic analysis (LSA) and correspondence analysis (CA). For the lowest level of dimension, precision of correspondence analysis (CA) method lies well above that obtained with latent semantic analysis (LSA). But for high dimensional space, precision of latent semantic analysis (LSA) method is above than that obtained with correspondence analysis (CA). Thus, latent semantic analysis (LSA) captures some structure in the data in high dimensional space which is obscured when correspondence analysis (CA) is used. Similarly correspondence analysis (CA) performed better in representing the structure of data in lower dimension than latent semantic analysis (LSA). Figure 5.6 and 5.7 shows precision-recall curves where precision is

plotted as a function of recall (from 0.1 to 0.9) for dimension 250 and 500 respectively. These data represent average data from the 30 queries available with the *MED* collection. These are typical precision-recall curves, with precision decreasing as recall increases. The important thing to notice is the difference between latent semantic analysis (LSA) and correspondence analysis (CA) methods. In 500-dimensional space latent semantic analysis (LSA) representation results in better performance in the discrimination of relevant from irrelevant documents. Similarly it can be said for correspondence analysis (CA) in 200-dimensional space.

V.5 Concluding Remarks

In this final chapter we have assessed the performance of correspondence analysis as compared to a method named, latent semantic analysis, which is especially useful for analyzing high dimensional sparse contingency table data. Our comparison concludes that correspondence analysis (CA) can be very useful method even for high dimensional data when the representation is sought on a smaller dimensional subspace.

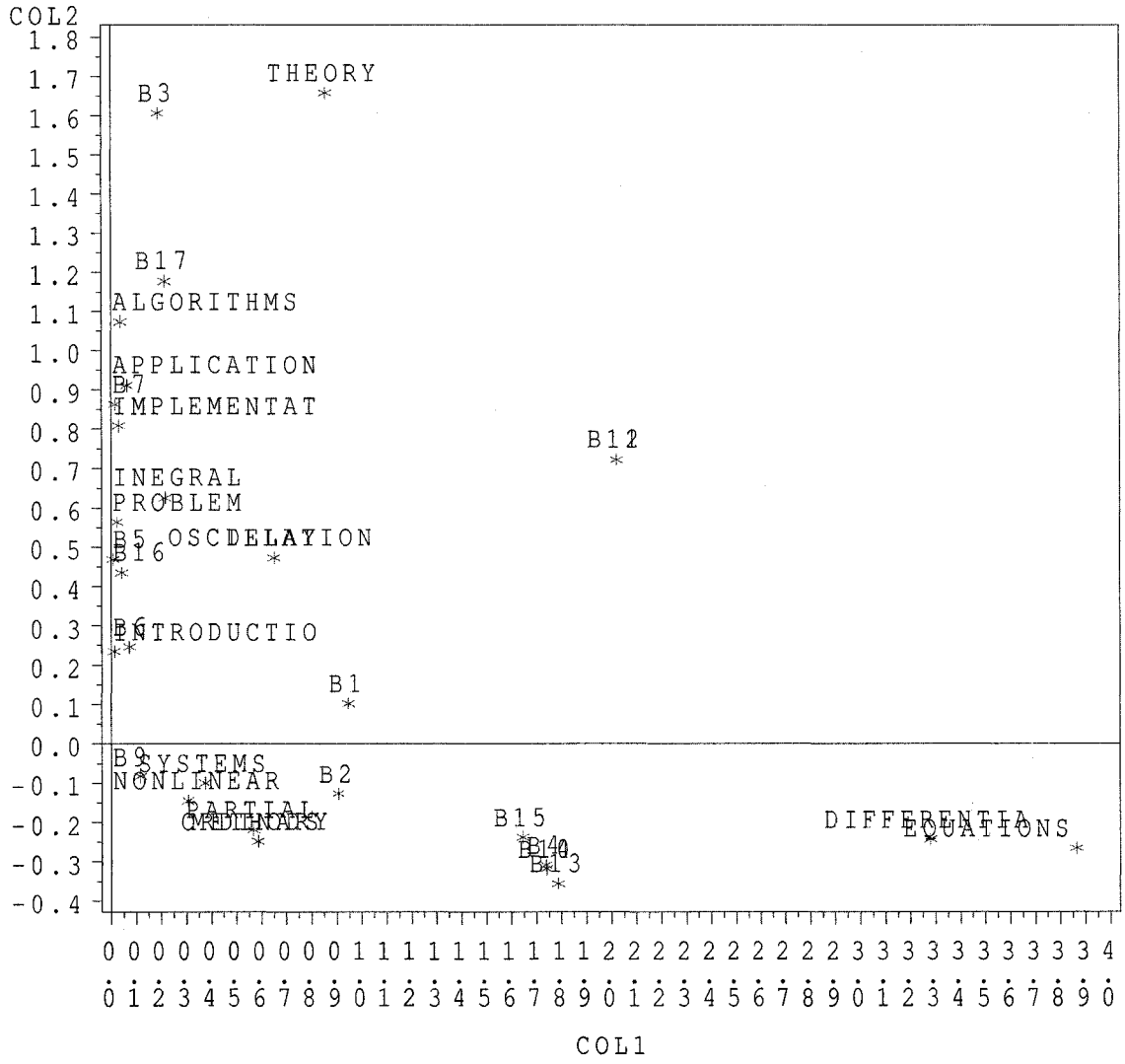

Figure 5.1: LSA: Two-Dimensional Plot of Terms and Documents.

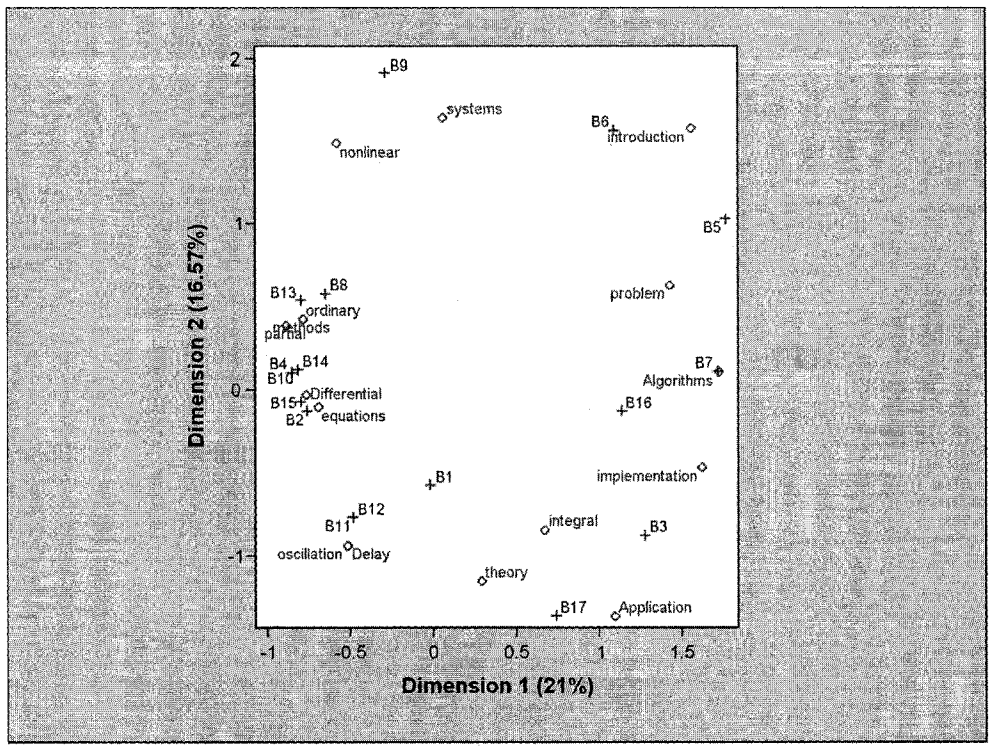

Figure 5.2: CA: Two-Dimensional Plot of Terms and Documents.

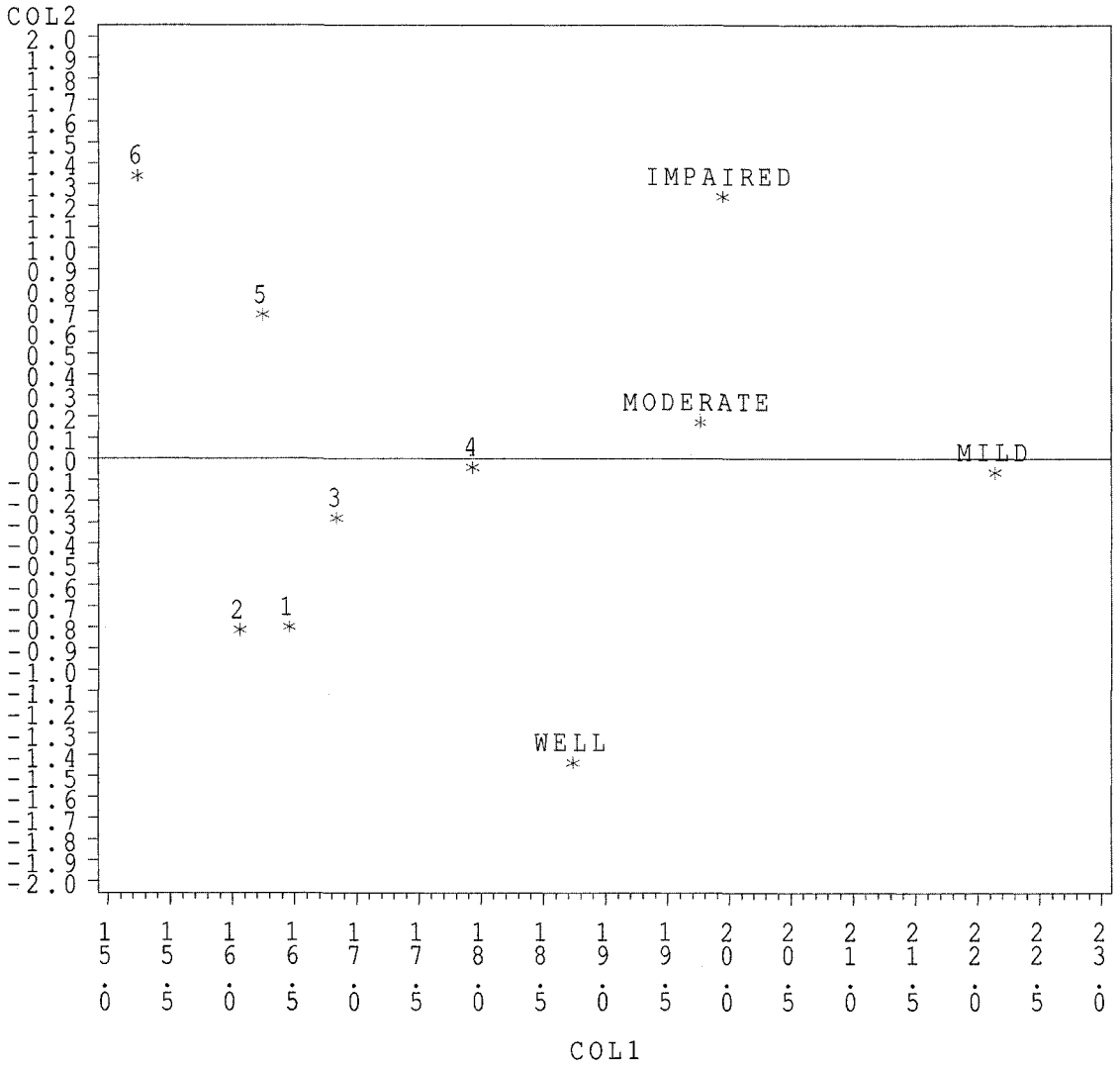

Figure 5.3: LSA: Two-Dimensional Plot of Socioeconomic Status by Mental Health of Children Data.

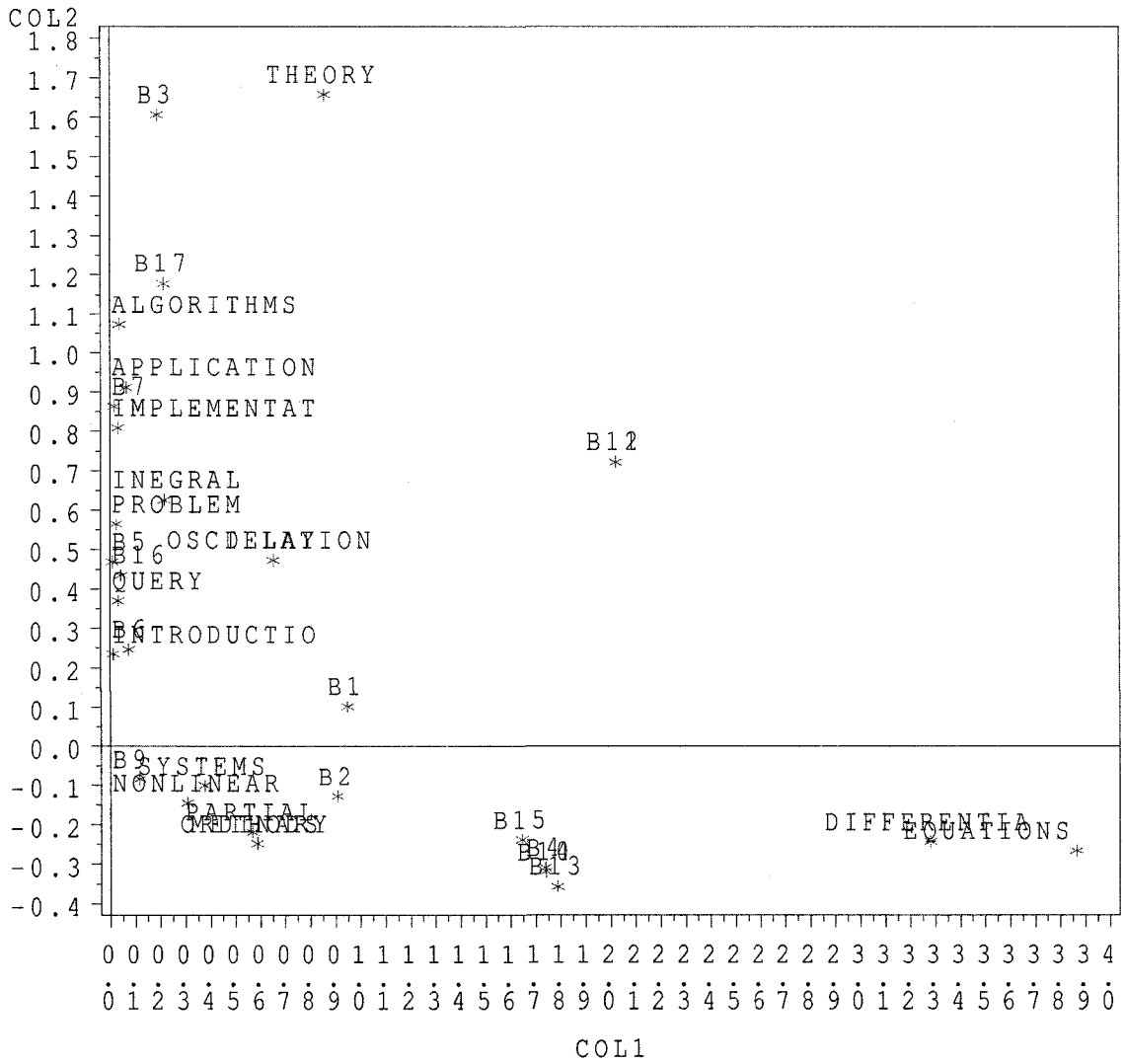

Figure 5.4: LSA: Two-Dimensional Plot of Query Vector.

Table 5.7: LSA: Ranked Documents Based on their Cosine

Document	Cosine
B16	0.9999227
B3	0.9995526
B5	0.9977227
B7	0.9977114
B17	0.9955983
B6	0.9788947
B11	0.4170113
B12	0.4170113
B1	0.192965
B2	-0.051833
B15	-0.056541
B4	-0.089106
B10	-0.094128
B14	-0.094128
B13	-0.109775
B8	-0.138082
B9	-0.508324

Table 5.8: CA: Ranked Documents Based on their Cosine

Document	Cosine
B17	0.9998468
B3	0.8982405
B1	0.8584114
B16	0.5861591
B11	0.478505
B12	0.478505
B7	0.4355108
B5	-0.01321
B2	-0.338412
B15	-0.402618
B6	-0.433028
B4	-0.595136
B10	-0.613468
B14	-0.613468
B13	-0.893458
B9	-0.933531
B8	-0.946973

Table 5.9: Characteristics of MED *Dataset*

Number of Documents	1033
Number of Indexing Terms	5478
Percentage of Nonzero entries in Matrix	0.91
Number of Queries	30
Number of Queries of relevant documents	696
Avg. No of Relevant Document per Query	23

Table 5.10: Average Precision of Correspondence Analysis (CA) for MED *Dataset*

Dimension	Documents Retrieved (Doc_t)										Avp_{qd}
	60	120	180	240	300	360	420	480	540	600	
10	0.60	0.61	0.59	0.56	0.54	0.51	0.51	0.50	0.49	0.48	0.54
20	0.72	0.70	0.69	0.65	0.63	0.63	0.60	0.60	0.58	0.56	0.64
30	0.77	0.75	0.72	0.70	0.69	0.66	0.65	0.64	0.62	0.61	0.68
40	0.77	0.76	0.76	0.76	0.74	0.71	0.70	0.68	0.66	0.65	0.72
50	0.78	0.79	0.76	0.74	0.75	0.73	0.72	0.69	0.67	0.65	0.73
60	0.83	0.80	0.78	0.73	0.72	0.72	0.71	0.69	0.67	0.66	0.73
70	0.78	0.78	0.77	0.74	0.74	0.73	0.72	0.70	0.68	0.65	0.73
80	0.80	0.80	0.76	0.74	0.73	0.70	0.70	0.68	0.66	0.64	0.72
90	0.80	0.80	0.77	0.72	0.71	0.70	0.68	0.66	0.64	0.63	0.71
100	0.82	0.75	0.73	0.71	0.72	0.70	0.68	0.65	0.63	0.62	0.70
150	0.80	0.75	0.71	0.67	0.66	0.65	0.64	0.63	0.61	0.59	0.67
200	0.70	0.68	0.67	0.64	0.63	0.63	0.60	0.57	0.56	0.54	0.62
250	0.70	0.67	0.64	0.62	0.60	0.58	0.57	0.56	0.54	0.52	0.60
300	0.68	0.66	0.61	0.59	0.56	0.54	0.53	0.51	0.49	0.48	0.57
350	0.63	0.58	0.59	0.57	0.54	0.51	0.50	0.48	0.46	0.46	0.53
400	0.62	0.58	0.56	0.53	0.53	0.50	0.48	0.48	0.47	0.45	0.52
450	0.58	0.58	0.55	0.52	0.51	0.49	0.47	0.46	0.45	0.43	0.50
500	0.58	0.57	0.54	0.52	0.49	0.48	0.46	0.45	0.44	0.43	0.50
550	0.62	0.58	0.54	0.50	0.50	0.48	0.46	0.44	0.43	0.41	0.49
600	0.60	0.54	0.53	0.51	0.49	0.46	0.45	0.43	0.42	0.40	0.48
650	0.58	0.53	0.51	0.50	0.48	0.46	0.44	0.42	0.40	0.38	0.47
700	0.60	0.55	0.50	0.49	0.47	0.46	0.43	0.41	0.39	0.38	0.47
750	0.55	0.53	0.51	0.47	0.46	0.44	0.41	0.39	0.38	0.37	0.45
800	0.55	0.53	0.47	0.44	0.44	0.43	0.40	0.38	0.37	0.35	0.44
850	0.57	0.51	0.46	0.44	0.43	0.41	0.40	0.37	0.36	0.35	0.43
900	0.57	0.51	0.46	0.45	0.41	0.40	0.40	0.37	0.35	0.34	0.42
950	0.52	0.49	0.46	0.43	0.39	0.40	0.38	0.36	0.35	0.34	0.41
1000	0.50	0.48	0.43	0.42	0.40	0.38	0.37	0.36	0.35	0.34	0.40

Table 5.11: Average Precision of Latent Semantic Analysis (LSA) for MED *Dataset*

Dimension	Documents Retrieved (Doc_t)										Avp_{qd}
	60	120	180	240	300	360	420	480	540	600	
10	0.37	0.28	0.28	0.28	0.28	0.28	0.26	0.25	0.24	0.24	0.28
20	0.48	0.44	0.43	0.41	0.40	0.40	0.39	0.39	0.38	0.37	0.41
30	0.50	0.47	0.47	0.46	0.46	0.44	0.42	0.41	0.40	0.40	0.44
40	0.52	0.50	0.48	0.50	0.49	0.48	0.46	0.46	0.44	0.43	0.48
50	0.53	0.55	0.53	0.50	0.52	0.51	0.49	0.47	0.46	0.44	0.50
60	0.55	0.53	0.55	0.54	0.50	0.49	0.48	0.47	0.45	0.44	0.50
70	0.62	0.61	0.55	0.55	0.51	0.49	0.49	0.47	0.46	0.45	0.52
80	0.60	0.61	0.56	0.54	0.52	0.51	0.49	0.48	0.47	0.46	0.52
90	0.63	0.61	0.57	0.58	0.57	0.55	0.52	0.50	0.47	0.46	0.55
100	0.62	0.63	0.60	0.55	0.53	0.53	0.51	0.50	0.48	0.47	0.54
150	0.67	0.64	0.63	0.58	0.57	0.55	0.52	0.50	0.49	0.48	0.56
200	0.77	0.68	0.62	0.59	0.57	0.57	0.56	0.53	0.51	0.48	0.59
250	0.77	0.68	0.63	0.60	0.57	0.56	0.53	0.52	0.50	0.49	0.59
300	0.78	0.69	0.63	0.59	0.56	0.56	0.53	0.51	0.48	0.47	0.58
350	0.78	0.73	0.63	0.60	0.57	0.55	0.51	0.50	0.48	0.46	0.58
400	0.77	0.68	0.63	0.61	0.55	0.53	0.51	0.50	0.48	0.45	0.57
450	0.78	0.68	0.65	0.59	0.55	0.52	0.50	0.49	0.47	0.45	0.57
500	0.82	0.68	0.64	0.60	0.55	0.53	0.49	0.49	0.47	0.45	0.57
550	0.80	0.68	0.65	0.59	0.55	0.53	0.50	0.46	0.46	0.44	0.56
600	0.82	0.66	0.63	0.58	0.53	0.52	0.48	0.46	0.44	0.43	0.56
650	0.77	0.65	0.62	0.59	0.54	0.51	0.48	0.45	0.43	0.42	0.55
700	0.78	0.68	0.62	0.58	0.55	0.50	0.48	0.45	0.44	0.42	0.55
750	0.78	0.67	0.59	0.56	0.52	0.48	0.47	0.45	0.43	0.40	0.53
800	0.78	0.64	0.58	0.53	0.50	0.48	0.46	0.44	0.41	0.40	0.52
850	0.75	0.62	0.55	0.53	0.50	0.48	0.44	0.42	0.40	0.39	0.51
900	0.73	0.60	0.57	0.52	0.48	0.48	0.45	0.42	0.40	0.39	0.50
950	0.72	0.62	0.54	0.52	0.49	0.46	0.42	0.40	0.39	0.39	0.49
1000	0.67	0.57	0.52	0.48	0.46	0.44	0.41	0.40	0.39	0.38	0.47

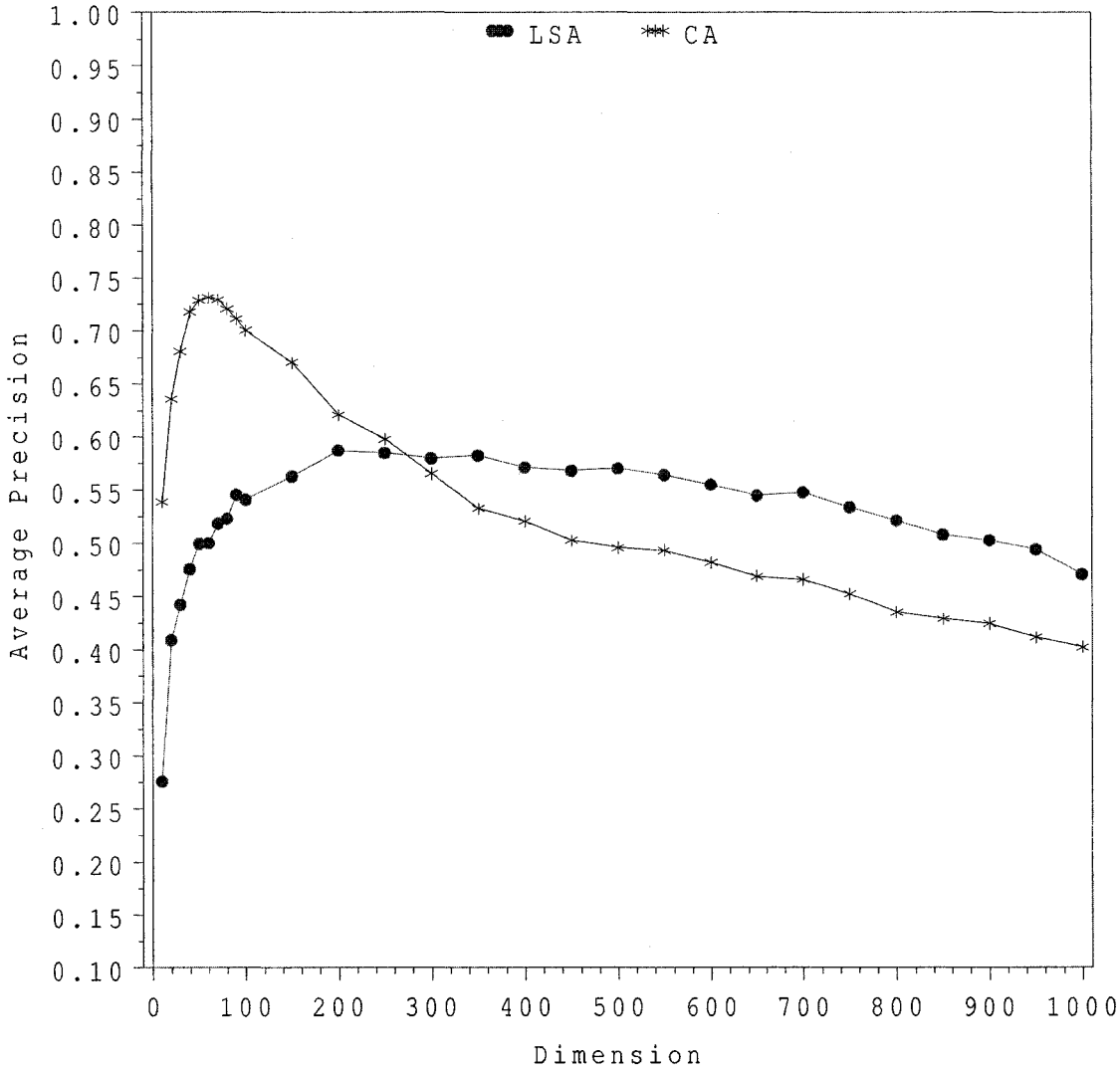

Figure 5.5: MED: Average Precision $(\mathbf{Avp}_{qd})$ *as a Function of Dimension.*

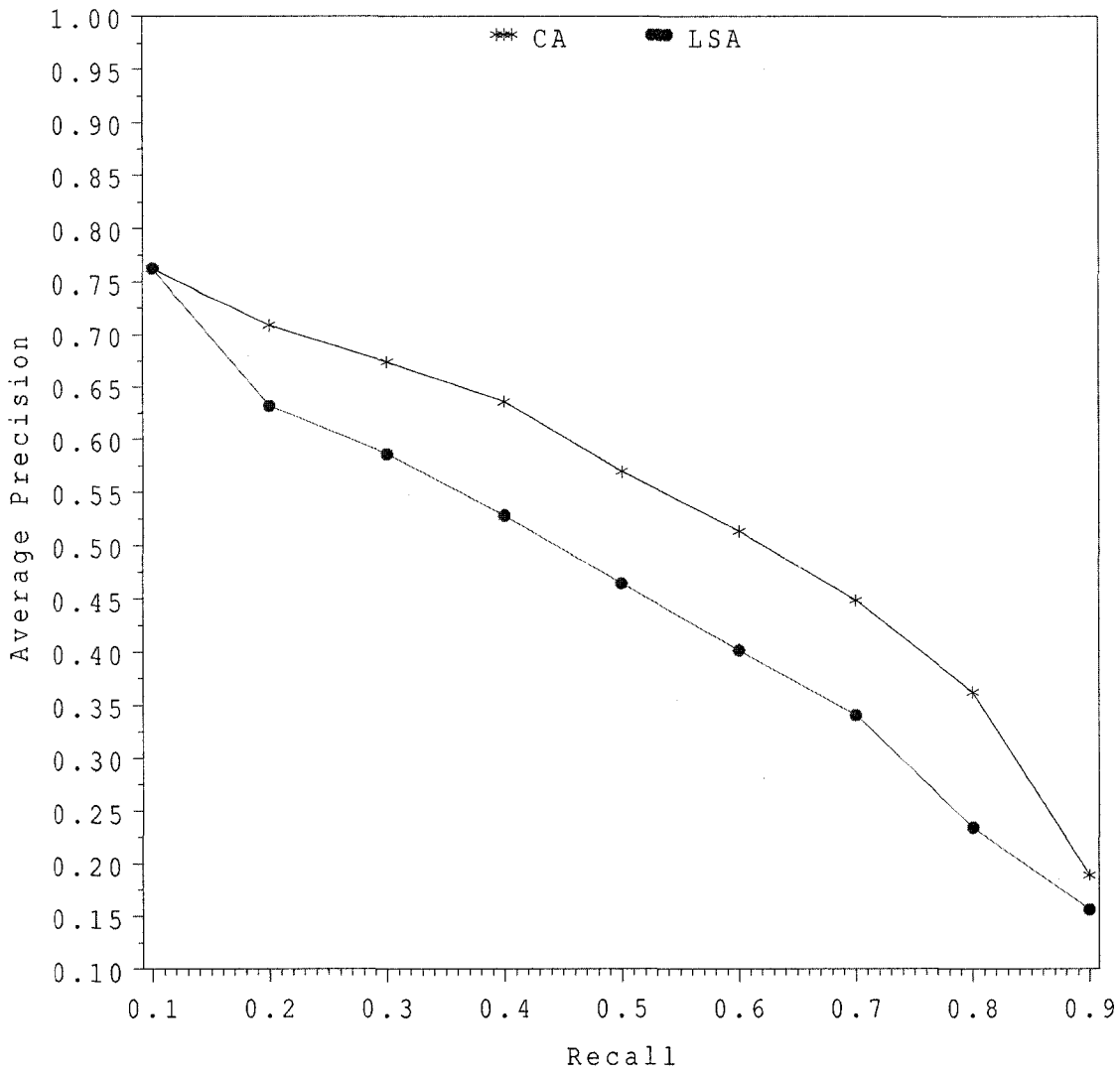

Figure 5.6: MED: Precision-Recall Curve for 200-Dimensional Space.

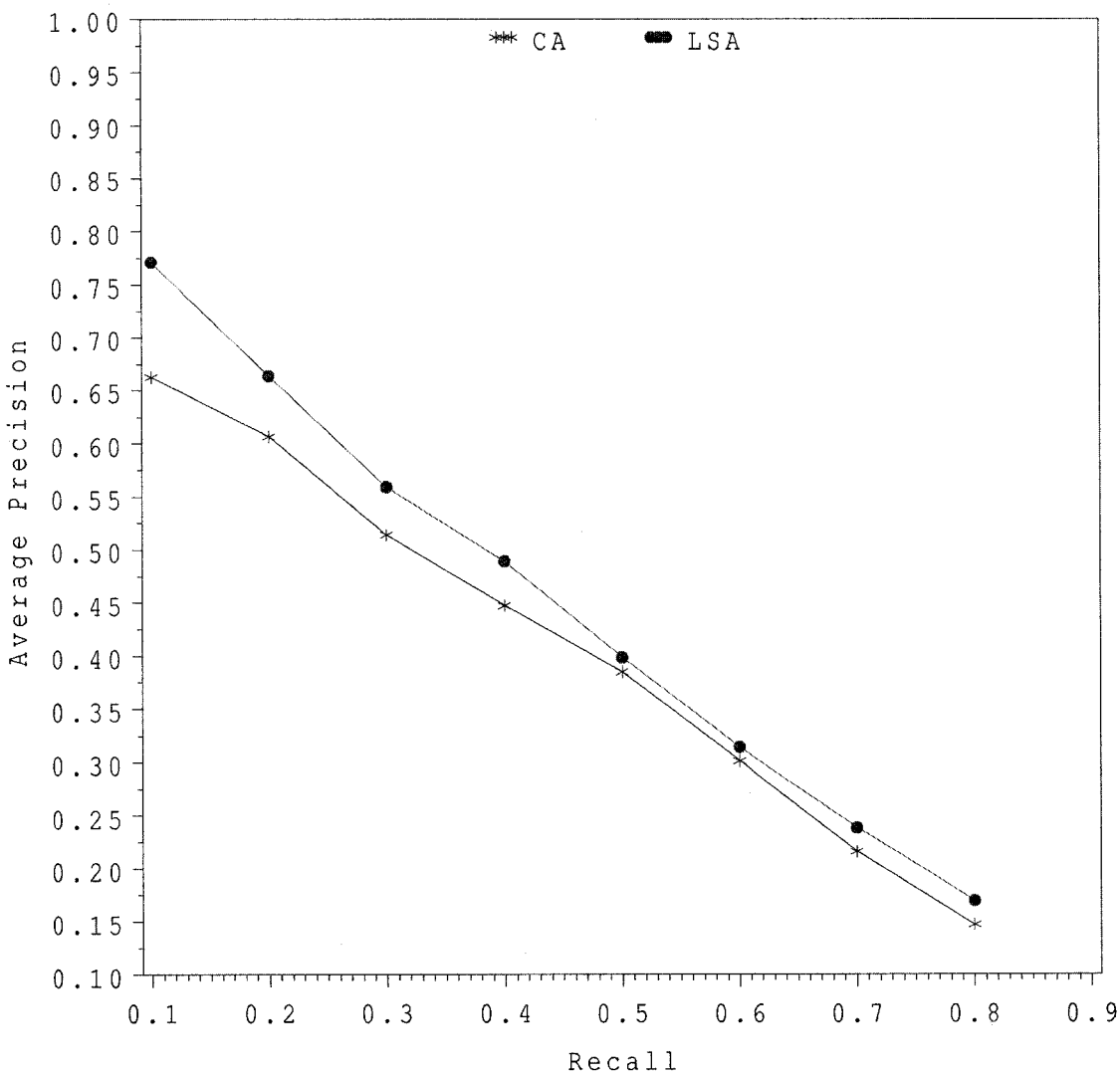

Figure 5.7: MED: Precision-Recall Curve for 500-Dimensional Space.

REFERENCES

Bartlett, M. S. (1937), "Properties of Sufficiency and Statistical Tests", *Proc. R. Soc. Lond*, 160, 268-282.

Beaghen, M. (1997), "Canonical Variate Analysis and Related Methods with Longitudinal Data", *Ph.D. thesis, Department of Statistics, Virginia Tech.*

Benzécri, J. P. (1969), "Statistical Analysis as a Tool to Make Patterns Emerge from Data", *In Methodologies of Pattern Recognition*, 35-60.

Benzécri, J. P. (1992), *Correspondence Analysis Handbook*, New York: Marcel Dekker.

Berry, M. W., and Browne, M. (2005), *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, Philadelphia: SIAM.

Box, G. E. P. (1949), "A General Distribution Theory for a Class of Likelihood Criteria", *Biometrika*, 36, 317-346.

Chaganty, N. R., and Mav, D. (2007), "Estimation methods for analyzing longitudinal data occurring in biomedical research",To appear in *Computational Methods in Biomedical Research.*

Chaganty, N. R., and Naik, D. N. (2002), "Analysis of Multivariate Longitudinal Data using Quasi-Least Squares", *Journal of Statistical Planning and Inference*, 103, 421-436.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990), "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, 41, 391-407.

Dumais, S. T. (1991), "Improving the Retrieval of Information from External Sources", *Behavior Research Methods, Instruments, & Computers*, 23, 229-236.

Gauch, H. G., Chase, G. B., and Whittaker, R. H. (1974), "Ordination of vegetation Samples by Gaussian Species Distributions.", *Ecology*,55,1382-1390.

Goodman, L. A., (1981), "Association Models and Canonical Correlations in the Analysis of Cross-Classification having Ordered Categories.", *Journal of American Statistical Association*,76, 320-334.

Greenacre, M. J. (1984), *Theory and Applications of Correspondence Analysis*, London: Academic Press.

Harman, D. (1992), *Ranking Algorithms, in Information Retrieval: Data Structures & Algorithms*, New Jersey:Prentice-Hall, pp. 363-392.

Hegde, L. M., and Naik, D. N. (1999), "Canonical Correspondence Analysis in SAS Software.", *Proceedings of the Twenty-Fourth Annual SAS Users Group International Conference*,1999, 1607-1613.

Hegde, L. M., and Naik, D. N. (2006), "Cannonical Correspondence Analysis: Some New Interpretations and Computations using SAS.", *Preprint.*

Hill, M. O., (1974), "Correspondence Analysis: A Neglected Multivariate Method", *Applied Statistics*, 23, 340-354.

Hotelling, H. (1936), "Relations Between Two Sets of Variates", *Biometrika*, 28, 321-377.

Johnson, R. A., and Wichern D. W. (2002), *Applied Multivariate Statistical Analysis*, New Jersey: Prentice-Hall.

Kettenring, J. R. (1971), "Canonical Analysis of Several Sets of Variables", *Biometrika*, 58, 433-451.

Khattree, R., and Naik, D. N. (2000), *Multivariate Data Reduction and Discrimination with SAS Software.*, North Carolina: Wiley-SAS.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998), "An Introduction to Latent Semantic Analysis", *Discourse Processes*, 25, 259-284.

Mardia, K. V., Kent, J. J., & Bibby, J. M. (1979), *Multivariate Analysis*, New York: Academic Press.

Mav, D., (2004), *Statistical Analysis of Longitudinal and Multivariate Discrete Data*, PhD dissertation, Department of Mathematics and Statistics, Old Dominion University.

Naik, D. N., and Rao, S. (2001), "Analysis of Multivariate Repeated Measures Data with a Kronecker Product Structured Covariance Matrix", *Journal of Applied Statistics*, 28, 91-105.

Olkin, I., and Tate, R.F. (1961), "Multivariate Correlation Models With Mixed Discrete and Continuous Variables", *Annals of Mathematical Statistics*, 32, 448-465 (correction in 36, 343-344).

O'Neill, M.E., (1981), "A note on the canonical correlations from contingency tables", *Australian Journal of Statistics*, 23, 58-66.

Pearson, E. S. (1969), "Some Comments on the Accuracy of Box's Approximations to the Distribution of M", *Biometrika*, 56, 219-220.

Salton G., and McGill, M. (1983), *Introduction to Modern Information Retrieval*, New York: McGraw-Hill.

Sim, C. H. (1993), "Generation of Poisson and Gamma Random Vectors With Given Marginals and Covariance Matrix," *Journal of Statistical Computation and Simulation*, 47, 1–10.

Srole, L., Langner, T. S., Michael, S. T., Kirkpatrick, P., Opler, M. K. and Rennie, T. A. C. (1978), *Mental Health in the Metroplois: The Midtown Manhattan Study*, New York: New York University Press.

Ter Braak, C. J. F. (1986), "Canonical Correspondence Analysis: A New Eigen Vector Technique for Multivariate Direct Gradient Analysis", *Ecology*,67,1167-1179.

Ter Braak, C. J. F. (1985), "Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model.", *Biometrics*,41,859-873.

Van der Aart, P. J. M. and Smeek-Enseink, N. (1975), "Correlations Between Distributions of Hunting Spiders and Environmental Characteristics in a Dune Area", *Netherlands Journal of Zoology*,25,1-45.

APPENDIX

MULTIVARIATE POISSON SIMULATIONS IN SAS

```
*SAS Subroutines for Multivariate Poisson Simulations;
/*-----------------------------------------------------------------*/
/* The following subroutine simulates nsims multivariate Poisson */
/* obs from given covariance matrix Sigma using Sim's algorithm */
/* as given in Deepak Mav PhD thesis (2004).  */
/*-----------------------------------------------------------------*/

START SIMPOI(seed, Sigma, nsims);
   RUN Decompose(Sigma, alpha, lambda, Error, m);
   if (Error < 0) then do;
      print "Simulations Failure";
      return(Error);
   end;

   Z = J(m, nsims, 0);
   do k = 1 to nsims;
      X = J(m, 1, 0);
      do j = 1 to m;
         do i = 1 to j-1;
            if(X[i] & (alpha[j,i] > 0)) then
               Z[j,k] = Z[j,k] +
                  RANBIN(seed, X[i], alpha[j,i]);
         end;
         X[j] = RANPOI(seed, lambda[j]);
         Z[j,k] = Z[j,k] + X[j];
      end;
   end;
   return(Z);
Finish SIMPOI;
```

```
Start Decompose(Sigma, alpha, lambda, Error, m);
   m = nrow(Sigma); alpha = I(m); lambda = J(m, 1, 0); Error=1;
   lambda[1] = Sigma[1,1];
   do j = 2 to m;
      alpha[j,1] = Sigma[1,j]/lambda[1];
      if((0 > alpha[j,1]) | (alpha[j,1] > 1)) then Error = -1;
      do i = 2 to (j-1);
         do k = 1 to (i-1);
            alpha[j,i] = alpha[j,i] +
               alpha[i,k]*alpha[j,k]*lambda[k];
         end;
         alpha[j,i] = (Sigma[i,j] - alpha[j,i])/lambda[i];
         if((0 > alpha[j,i]) | (alpha[j,i] > 1)) then Error = -1;
      end;
      do k = 1 to (j-1);
         lambda[j] = lambda[j] + alpha[j,k]*lambda[k];
      end;
      lambda[j] = Sigma[j,j] - lambda[j];
      if(lambda[j] <= 0) then Error = -2;
   end;
Finish Decompose;


/*-----------------------------------------------------------------*/
/* This subroutine computes first four central moments of Poisson */
/* random variables.The functional arguments are alpha and lambda */
/*-----------------------------------------------------------------*/
Start Moments(alpha, lambda);
   m = nrow(lambda); dim = m+m*(m+1)/2; V = J(dim, dim, .);
   /* Second order moments */
   do i = 1 to m;
      do j = 1 to i;
         value = alpha[Unique(i||j),]; value =value[#,]*lambda;
         V[i,j] = value; V[j,i] = value;
      end;
```

```
   end;
   /* Third order moments */
   do i = 1 to m;
      do j = 1 to i;
         index1 = m + i*(i-1)/2+j;
         do k = 1 to m;
            value = alpha[Unique(i||j||k),];
            value = value[#,]*lambda;
            V[index1, k] = value; V[k, index1] = value;
         end;
      end;
   end;
   /* Fourth order moments */
   do i = 1 to m;
      do j = 1 to i;
         index1 = m + i*(i-1)/2+j;
         do k = 1 to m;
            do l = 1 to k;
               index2 = m + k*(k-1)/2+l;
               if (index1 >= index2) then do;
                  value = alpha[Unique(i||j||k||l),];
                  value = value[#,]*lambda;
                  V[index1,index2] = value + V[i,k]*V[j,l]
                                   + V[i,l]*V[j,k];
                  V[index2,index1] = V[index1,index2];
               end;
            end;
         end;
      end;
   end;
   return(V);
Finish Moments;
```

VITA

Jayesh Srivastava
Department of Computational and Applied Mathematics
Old Dominion University
Norfolk, VA 23529

Education

Ph.D.	Old Dominion University, Norfolk, VA. (May 2007)
	Major: Computational and Applied Mathematics (Statistics)
MS	Indian Institute of Technology, Mumbai, India. (August 1998)
	Major: Mathematics,
BS	University of Nagpur, Nagpur, India. (June 1996)

Experience

Statistics Graduate Assistant (08/2005 - 05/2007)
Old Dominion University, Norfolk, VA

Teaching Assistant (08/2002 - 08/2005)
Old Dominion University, Norfolk, VA

Typeset using LaTeX.